



中国研究生创新实践系列大赛

“华为杯”第十七届中国研究生

数学建模竞赛

学　　校	南京林业大学
参赛队号	20102980116
队员姓名	1.曹磊
	2.王羽尘
	3.王玉鹏

中国研究生创新实践系列大赛

“华为杯”第十七届中国研究生

数学建模竞赛

题 目 降低汽油精制过程中的辛烷值损失模型

摘 要：

辛烷值是反映汽油燃烧性能的重要指标，现有技术在对催化裂化汽油进行脱硫和降烯烃过程中，普遍降低了汽油辛烷值，造成了辛烷值（RON）的损失。为在汽油精制过程中进行优化操作，我们通过对其样本信息进行数据挖掘与模型建立，提供有价值的信息支持。

本文基于汽油精制过程中的样本信息数据，根据操作变量之间的高度非线性和相互强耦联性，进行数据挖掘与分析。同时，根据数据宏观层面的规律性、微观层面的差异性，建立具有针对性的多个数学模型，并使用 R^2，MSE，RMSE 等多种参数、模型比对与平稳性分析，进行了评估与合理性验证。本文所做的工作可概括为以下几点:

问题一：在判断数据合理性基础上，对样本数据进行预处理。首先，对照变量取值和 Δ 值范围，查找异常数据；其次，根据最大最小的限幅，剔除限幅外数据；然后，根据拉依达准则，去除异常值，共清洗剔除了附件三中的 726 条数据。最后，进行样本值确定，分别将 285 号和 313 样本操作变量的平均值与同一行进行比较，285 号的契合度达到 100%，313 号存在些许差，多数数值的契合度达到 90%以上。

问题二：根据数据特性，建立了 mRMR-IFS-SVR 模型。首先使用最大相关—最小冗余特征选择方法（mRMR）求解特征与目标损失值的相关性列表，获取 30 个主要变量的候选列表；其次，考虑到变量之间的高度非线性和相互强耦联，我们联用增量特征选择方法(IFS)，建立了高斯核函数支持量回归模型(rbf-SVR)，以评估参数筛选出了 24 个主要变量。同时，通过相关系数、互信息和最大信息数、随机森林、递归特征消除等方法验证主变量筛选的合理性。

问题三：本题主要通过数据挖掘技术建立辛烷值损失预测模型，即通过样本数据，求解最佳回归模型。由于数据本身的变量之间具有高度非线性和相互强耦联的关系，因此，本题选取相应的回归模型进行求解。我们主要联用第二问的 rbf-SVR 模型，与决策树回归、随机森林回归、梯度上升回归树进行比较评估参数选取最佳预测模型。最终确定了拟合优度为 93.75%的决策树回归模型。

问题四：基于合作博弈论，对筛选样本数据进行分析，并建立多目标优化模型。沿用决策树回归建立硫含量与 24 个主变量的预测模型，我们重点建立了“合作博弈——ParetTo 强度进化算法”：输入样本数据的操作变量，约束其上下限，根据“合作博弈”方法、“ParetTo 强度进化算法”，进行多次实验的多次迭代求近似最优解；获取多次迭代的最优解对应的

操作变量值取平均增降幅度，使用分布密度函数求得均值以及置信区间，提供整体操作变量的优化方案。同时，通过验证样本契合度、模型迭代的稳定性说明其合理性。

问题五：本题沿用了第四问中第 133 号样本的多次迭代最优解平均值作为增降服务的参考，使用贪心算法对主要变量进行逐步调整，总调整步长为 259，展现了汽油辛烷值和硫含量的变化轨迹。第 234 步时达到最优，辛烷值最高为 88.7，硫含量为 3.4 μg/g。当调整步长为 228 时，硫含量最低为 3.2 μg/g。

关键词：辛烷值；汽油精制；mRMR-IFS-SVR 模型；决策树回归；合作博弈论；PareTo 强度进化算法

目录

一、问题重述

问题一：数据处理

请参考近 4 年的工业数据(见附件一“325 个数据样本数据.xlsx”)的预处理结果，依“样本确定方法”（附件二）对 285 号和 313 号数据样本进行预处理（原始数据见附件三“285 号和 313 号样本原始数据.xlsx”）并将处理后的数据分别加入到附件一中相应的样本号中，供下面研究使用。

问题二：寻找建模主要变量

由于催化裂化汽油精制过程是连续的，所以实际情况中，可以通过操作过程降低辛烷值损失。同时，原料性质、待生吸附剂性质、再生吸附剂性质也可能对辛烷值有所影响。因此，通过筛选影响辛烷值降低的主要变量，能够进一步推进预测辛烷值模型的建立，为后续的汽油实际精制奠定基础。

本题主要根据提供的 325 个样本数据（见附件一），通过降维的方法从 367 个操作变量中筛选出建模主要变量，使之尽可能具有代表性、独立性（为了工程应用方便，建议降维后的主要变量在 30 个以下），并请详细说明建模主要变量的筛选过程及其合理性。（提示：请考虑将原料的辛烷值作为建模变量之一）。

问题三：建立辛烷值（RON）损失预测模型

采用上述样本和建模主要变量，通过数据挖掘技术建立辛烷值（RON）损失预测模型，并进行模型验证。

问题四：主要变量操作方案的优化

现有技术在对催化裂化汽油进行脱硫和降烯烃过程中，普遍降低了汽油辛烷值，造成了辛烷值（RON）的损失。由此可知，降低产品硫含量和减少辛烷损失为存在冲突。本研究主要目的在于保证产品硫含量不大于 5 μg/g 的前提下，分析辛烷值（RON）损失降幅大于 30%的样本对应的主要变量优化后的操作条件。考虑到优化过程中原料、待生吸附剂、再生吸附剂的性质保持不变，因此以它们在样本中的数据为准。

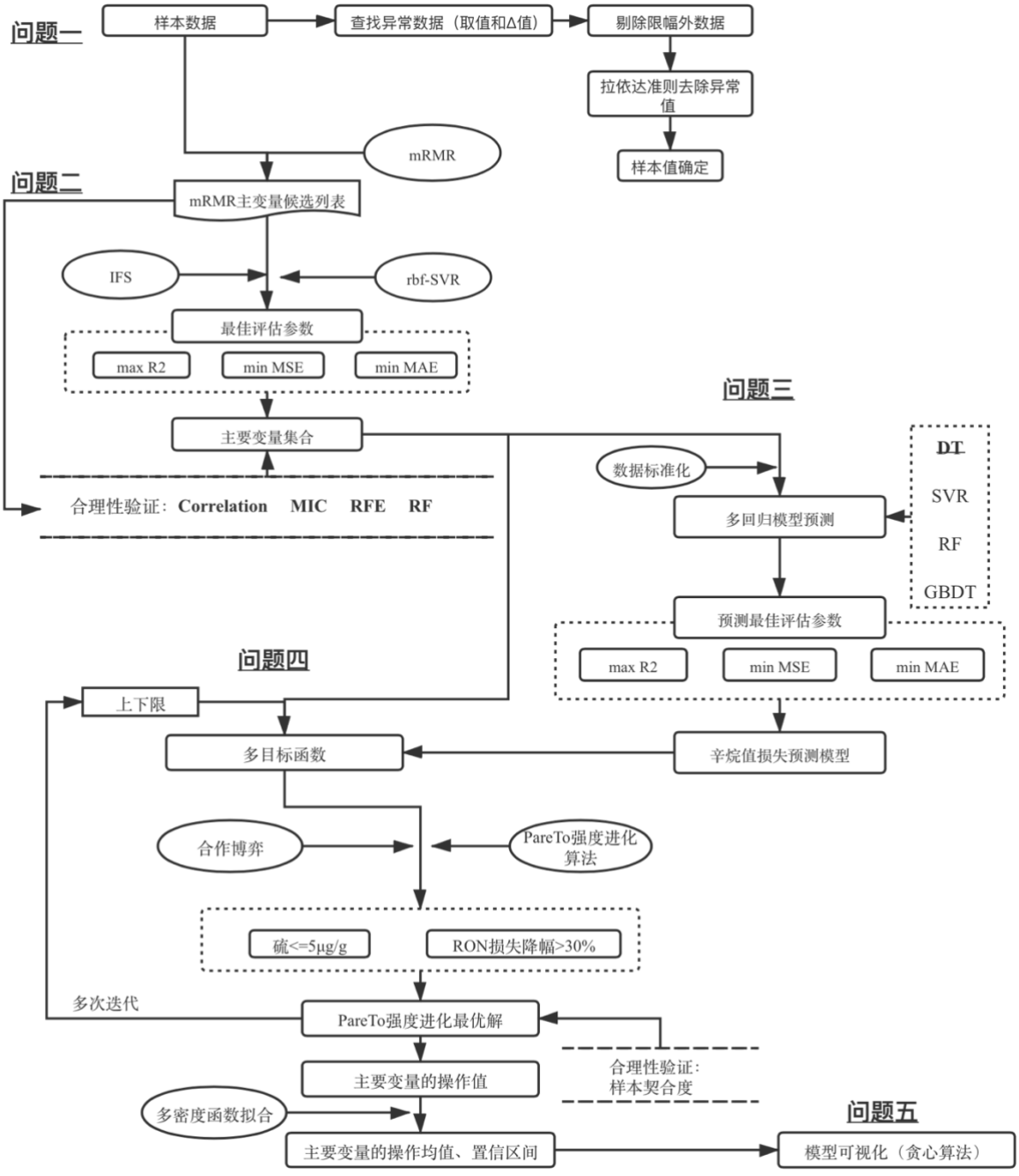
问题五：模型的可视化展示

工业装置为了平稳生产，优化后的主要操作变量（即：问题 2 中的主要变量）往往只能逐步调整到位，请你们对 133 号样本（原料性质、待生吸附剂和再生吸附剂的性质数据保持不变，以样本中的数据为准），以图形展示其主要操作变量优化调整过程中对应的汽油辛烷值和硫含量的变化轨迹。（各主要操作变量每次允许调整幅度值 Δ 见附件四“354 个操作变量信息.xlsx”）。

二、基本假设

1.假定数据采集为某石化企业的连续四年的真实数据；
2.假定数据样本的各个变量数值都在规定的数据范围之内；
3.假定辛烷值损失除多个变量外，不受外界影响；
4.假设产品性质无法改变，多数性质互不影响；
5.假定原料辛烷值为主要变量之一；
6.假设优化过程中原料、待生吸附剂、再生吸附剂的性质保持不变；
7.假设操作变量之间具有高度的非线性和相互强耦联性；
8.假定为平稳生产，工业装置及优化后的主要操作变量需逐步调整。

三、本题研究流程


问题一
样本数据
查找异常数据（取值和Δ值）
剔除限幅外数据
拉依达准则去除异常值
样本值确定
mRMR
问题二
mRMR主变量候选列表
IFS
rbf-SVR
最佳评估参数
max R2
min MSE
min MAE
主要变量集合
合理性验证：Correlation MIC RFE RF
问题三
数据标准化
多回归模型预测
DT
SVR
RF
GBDT
预测最佳评估参数
max R2
min MSE
min MAE
辛烷值损失预测模型
问题四
上下限
多目标函数
合作博弈
ParеTo强度进化算法
硫<=5μg/g
RON损失降幅>30%
多次迭代
ParеTo强度进化最优解
合理性验证：样本契合度
主要变量的操作值
多密度函数拟合
主要变量的操作均值、置信区间
问题五
模型可视化（贪心算法）

四、数据处理过程与分析

4.1 数据分析与预处理

本题所给出的附件三数据，是针对 285 号和 313 号样本的 354 个操作变量统计的。样本数据分别从 2017.07.17.06.03.00-08.00 和 2017.05.15.06.03.00-08.00，每隔 3 分钟采集的。本题主要分析流程图如下：

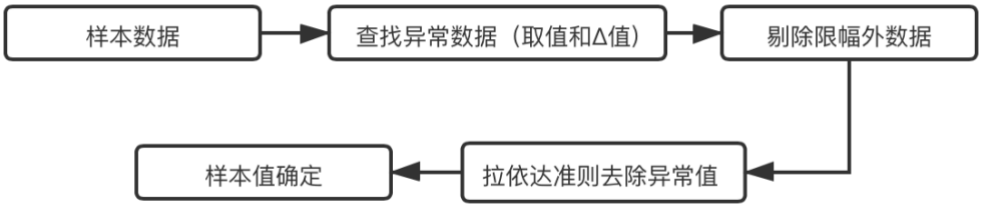


图 4.1 数据分析与处理流程

我们在判断数据合理性基础上，对样本数据进行预处理，主要分为以下三步：

（1）对照变量取值和 Δ 值范围，查找异常数据

根据附件四的操作变量信息，分别从取值和 Δ 值两个角度，查找范围之外的异常数据。通过编写代码将附件 3 与附件 4 的数据进行匹配，可知以下数据存在异常：

表 4.1 范围外异常数据

样本	序号	时间范围	变量编号	变量名称	取值范围	Δ 值
285	4-43	2017.07.17 06.03.00-08.00	49	原料缓冲罐液位	40-80	10
			84	再生烟气氧含量	0.5-3	0.2
			111	S-ZORB.FT_1204.TOTAL	45000-2500000	10000
313	45-84	2017.05.15. 06.03.00-08.00	22	精制汽油出装置硫含量	0-5	1
			49	原料缓冲罐液位	40-80	10
			84	再生烟气氧含量	0.5-3	0.2
			111	S-ZORB.FT_1204.TOTAL	45000-2500000	10000

将附件三中变量编号为 22、49、84、111 的整列数据与附件一中同年的数据进行对照，**两者数量级保持一致**，且**符合拉依达准则**（3 σ 准则），**因此予以保留。**通过查阅资料可知，其潜在原因是：①换算过程中存在单位转换，导致取值范围的变化，例如再生烟气氧含量是指进入再生系统的氧气含量，该值的单位随通风量而发生转化；② Δ 值的选取时间存在不统一的情况，例如 2017 年 4 月至 2019 年 9 月，数据采集频次为 3 分钟/次；2019 年 10 月至 2020 年 5 月，数据采集频次为 6 分钟/次。

（2）根据最大最小的限幅，剔除限幅外数据

结合附件四取值范围和附件 2 数据预处理的原则，分析 325 组样本和附件 3 数据的最大最小限幅范围，并以此剔除附件 3 里一部分不在此范围内的样本。针对 285 和 313 号样本，编写 Python 程序分别处理 40×354 矩阵。

分别基于 325 组样本和附件 3 数据的最大最小限幅原则筛选出 664 和 140 条限幅外的数据，并将筛选的数据进行对比：140 条限幅外的数据基本包含在 664 条数据里，仅有两条（67，323）和（69，323）未包含。因此根据最大最小的限幅原则，共剔除 666 条数据。

从图 4.2 所示中可以看出，样本 285 号所有的数值均在最大最小的限幅范围，样本 313 号存在较多超过最大最小的限幅范围的数值，且较为集中在 06:18:00-06:39:00 和 07:12:00-07:27:00，变量编号集中在 40-90、130-323 之间，需要针对这些点进行剔除。

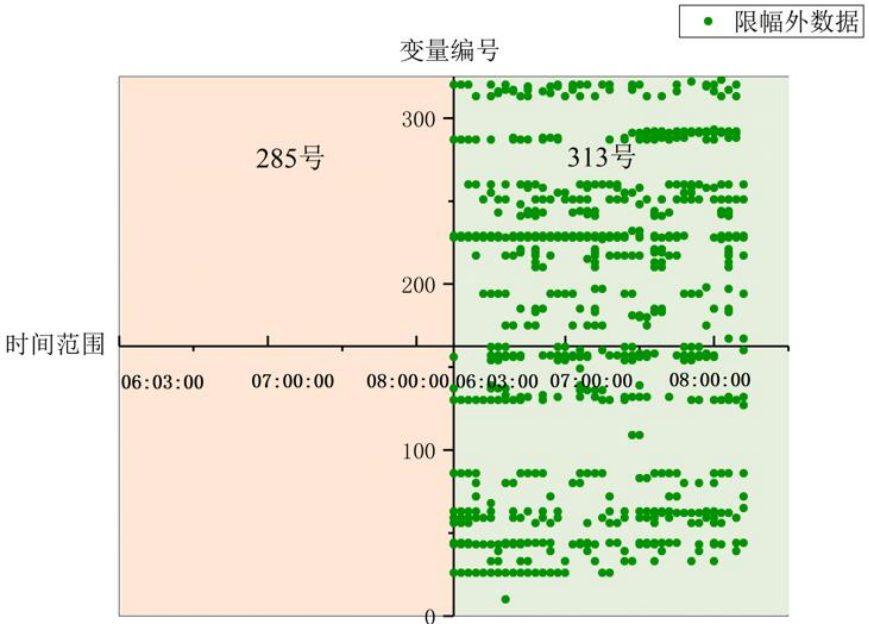


图 4.2　限幅外数据分布图

（3）拉依达准则（3 σ 准则），去除异常值

剔除限幅外数据后，依据拉依达准则（3 σ 准则）对附件 3 的操作变量剩下数据进行清洗，以去除异常值。通过 python 编程，去除含有粗大误差的异常样本：

$$\sigma=[\frac{1}{n-1}\sum_{i=1}^{n}v_i^2]^{1/2}=\{[\sum_{i=1}^{n}x_i{}^2-(\sum_{i=1}^{n}x_i)^2/n]/(n-1)\}^{1/2} \qquad (4\text{-}1)$$

通过拉依达准则清洗出 60 条异常数据，如图 3.2 所示。

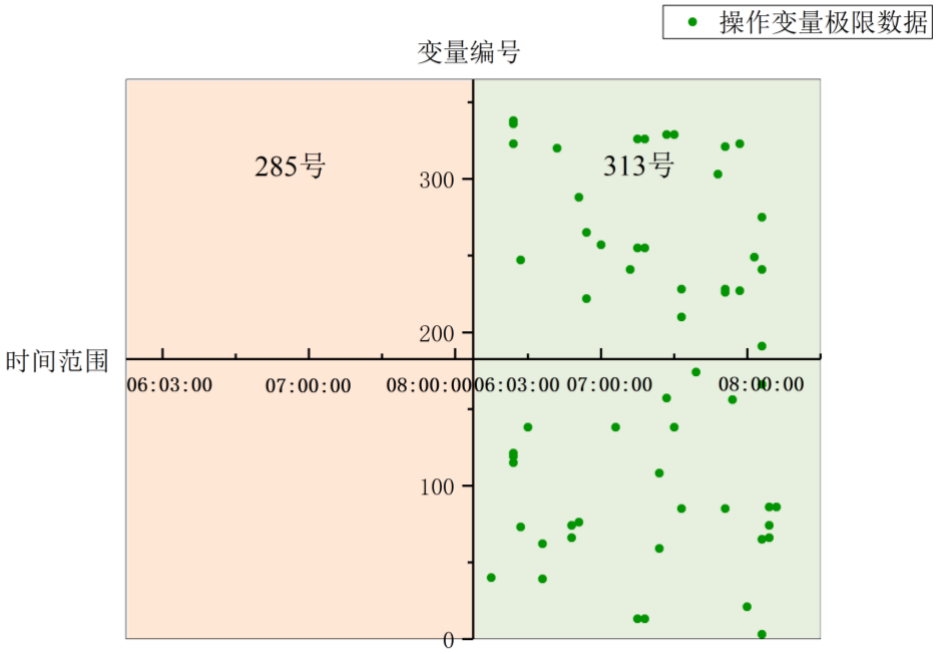


图 4.3 操作变量极限值分布图

从图 4.3 中可以看出，样本 285 号所有的数值均满足拉依达准则，样本 313 号存在部分不满足的数值，且数值较为分散，将这些数值予以删除。

同时，利用拉依达准则（3 σ 准则）对非操作变量进行清洗，其中有 37 条数据含有粗

大误差，结果如下。

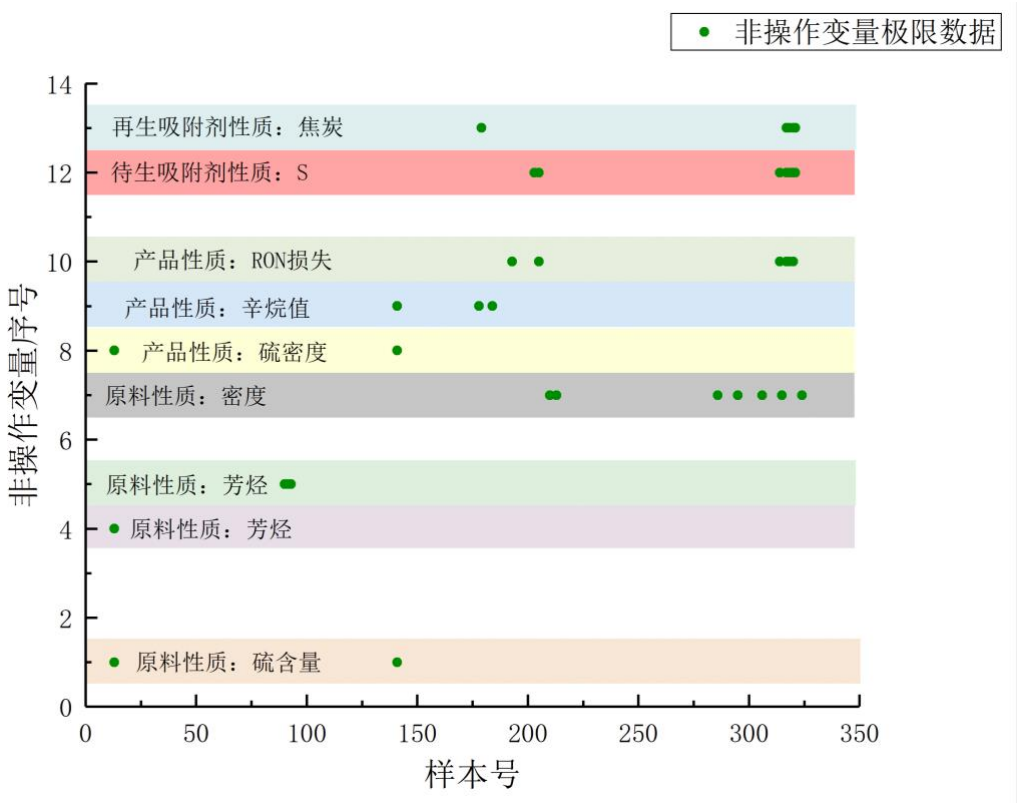


图 4.4 非操作变量极限值分布图

从图 4.4 中可以看出，14 组非操作变量中原料性质、产品性质和再生待生吸附剂性质中均有少数的异常数据，数值较为分散，集中在 140-215、280-324。

4.2 样本确定

由于辛烷值的测定数据相对于操作变量数据而言相对较少，本文参考附件 2，以辛烷值数据测定的时间点为基准时间，取其前 2 个小时的操作变量数据的平均值作为对应辛烷值的操作变量数据，最终结果如下：

（1）285 号样本操作变量

将求得的数据与附件 1 同一行数据进行比较，契合度均达到 100%，具体数值详见附录。

（2）313 号样本操作变量

将求得的数据与附件 1 同一行数据进行比较，对契合度非 100%的数据整理如表所示。

表 4.2 非契合数据整理表(部分表，详见附录)

变量编号	清洗后数据	附件 1 数据	契合度
…	…	…	100%
…	…	…	…
33	0.9932	0.9929	99.97%
…	…	…	…
…	…	…	…
43	**23.8203**	**44.4755**	**53.56%**
162	**18.8566**	**33.0150**	**57.12%**
228	**6.7575**	**0.9752**	**492.93%**

从表可知，313 号样本平均值与附件 1 数据存在些许差别，共 81 条数据的契合度未达

到 100%，其中变量编号 43、162、228 的契合度较低，其余数值的契合度相对较高，多数数值的契合度达到 90%以上，可能的原因是：

①附件 3 中变量编号 43 的样本数据最大值为 54.3626，共 18 个数值远高于最大值，极大数值比例达到 45%，根据样本确定方法将其剔除，导致求得的平均值小于附件 1 中给定的对应值。

②附件 3 中变量编号 162 的样本数据最大值为 69.5433，共 10 个数值大于 69.5433，同时存在 16 个数值为 0 的情况，剔除的 10 个极大数值对平均值影响很大，导致所求的平均值较低。

③附件 3 中变量编号 228 的样本数据最小值为-0.17175，有 33 个数值小于-0.17175，极小值比例达到 57.5%，根据样本确定方法将其剔除，只保留了部分较大的数值，导致求出的平均值远大于附件一中给出的值。

4.3 小结与讨论

（1）本题的数据来源为石化企业在四年中收集的历史数据，涉及到附件 1 和附件 3 的非操作变量和操作变量数据。

（2）从变量取值和 Δ 值范围、最大最小的限幅和拉依达准则三个方面进行数据预处理，取其前 2 个小时的操作变量数据的平均值作为对应辛烷值的操作变量数据，经过清洗后，共剔除 726 条数据。

（3）对比清洗后数据和附件 1 数据契合度可知，前者由于剔除了极大和极小值，导致求得的平均值与附件一中给出的值，存在部分偏差，有利于在实际情况中对工业装置信息采集数据的方式进行检测与调整。

五、主要变量的筛选与合理性验证

5.1 符号约定

符号	含义
D	数据样本集合
R	系数
$f(x)$	预测输出

5.2 问题分析

在建立辛烷值损失的工程应用中，需要对主要因素进行筛选，以便于进行后续的实际应用操作。本题采用通过先降维后建模的方法，研究影响辛烷损失的主要因素，可以忽略次要因素，使模型更加精确、独立，方便工程应用。基于不同变量的定义和作用，将 7 个原料性质、2 个待生吸附剂性质、2 个再生吸附剂性质、2 个产品性质等变量以及另外 354 个操作变量分为四类因素进行分析。

由于本题中，产品性质的 2 个因素（硫含量 μg/g, 辛烷值 RON）为结果因素，无法进行操作，则不纳入考虑范围内。

辛烷值损失=原料辛烷值-产品辛烷值

因此，本题需解决问题可分为以下两个重点：

1.操作变量（控制变量）之间具有高度非线性和相互强耦联的关系，因此选用有效应对高度非线性和相互强耦联的变量的方法进行降维。

2.根据提供的 325 个样本数据，从 365 个操作变量中筛选出建模主要变量，变量分类大致可分为以下 4 种。

（1）原料因素

原料性质对辛烷损失影响很大，原料所含的烷烃会使汽油辛烷值降低,，环烷烃及芳烃能使辛烷值增加。而硫含量、溴值和密度同样对辛烷值产生影响。

（2）吸附剂因素

待生吸附剂是指在吸附反应器系统中经过脱硫反应后得到的吸附剂，而再生吸附剂是指脱除炭和部分硫，由氢气还原表面的镍组分，活性得到恢复的吸附剂。随着吸附剂上残留的焦炭增加，降低了活性中心数目，导致转化率减少，从而减弱氢转移的反应，导致汽油辛烷值越高。

（3）操作因素

本研究涉及 354 个操作变量，在催化裂化汽油精制过程中，不同因素对辛烷损失也不同。例如，提高反应器温度使催化裂化反应速度加快，辛烷值上升。

（4）其他

经过催化裂化汽油精制装置得到的产品所具有的性质，不会对制作过程中辛烷的损失值造成影响，因此不考虑第 8、9、10 列的变量作为模型的因素（3 个产品信息变量）。

5.3 主要变量筛选的模型建立

在汽油精制过程中，辛烷值损失受到多方面因素的影响，包括将 7 个原料性质、2 个待生吸附剂性质、2 个再生吸附剂性质、2 个产品性质等变量以及另外 354 个操作变量。因此，建立降低辛烷值损失模型就需要降维以便于工程应用的研究。在进行筛选主要变量建模的过程中，我们着重考虑了炼油工业过程的复杂性以及设备的多样性，并且操作变量之间具有高度非线性和相互强耦联的关系。

本题的建模及求解的研究流程如图所示，首先使用最大相关—最小冗余特征选择方法

(Max-Relevance and Min-Redundancy, mRMR）[1]求解特征与目标损失值的相关性列表，获取 30 个主要变量的候选列表；其次，考虑到变量之间的高度非线性和相互强耦联，我们使用增量特征选择方法(Incremental Feature Selection, IFS)[2]，并建立了高斯核函数支持量回归模型(Gaussian kernel-Support Vector Regression, rbf-SVR)[3]，以拟合优度[4]、均方误差[5]、平均绝对误差[5]来筛选主要变量的最优子集。

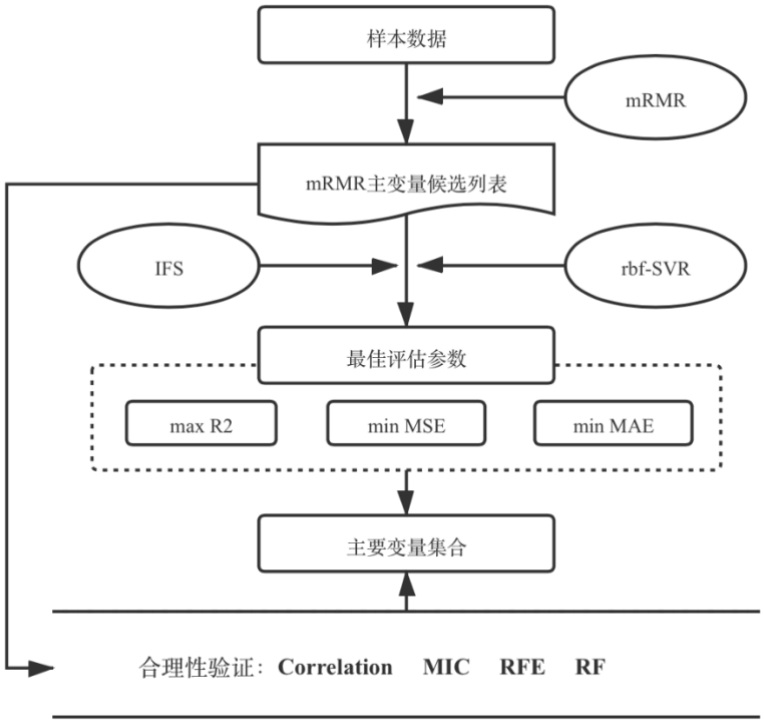


图 5.1 主要变量筛选的研究流程

5.3.1 最大相关—最小冗余特征选择（mRMR）

为方便工程应用，我们主要通过最大相关—最小冗余特征选择来确定主要变量候选集，以便进一步变量筛选。主要使用互信息和相关分数选择变量，并保证在存在其他所选变量的情况下，通过冗余项惩罚其相关性。设定数据样本为$S=(X,Y)$, 共有 m 个样本数据。其中，X 为影响辛烷值损失模型的变量集合，共有 n 维数据，Y 为辛烷值的产品结果集合。变量与辛烷目标值的最大依赖性公式为

$$\max D(X,Y),\quad D=I\left(x_i,i\in[1,n]\right) \tag{5-1}$$

即$n=1$，最大化互信息为$I\left(x_j,y\right),1\le j\le m$. 当$n>1$时，候选集中每次递增一个变量，则给定$n$-1个变量子集$X_{n\text{-}1}$集,当第 n 个变量选择，对于其互信息增加最大$I(X,Y)$的变量，其公式为

$$I(X_n;y)=\iint p(X_n,y)\log\frac{p(X_n,y)}{p(X_n)\,p(y)}dX_n d_y$$

$$=\iint p(X_{n\text{-}1},\mathrm{x}_n,y)\log\frac{p(X_{n\text{-}1},\mathrm{x}_n,y)}{p(X_{n-1},x_n)\,p(y)}dX_{n-1}dx_n d_y \quad (5\text{-}2)$$

$$=\int\cdots\int p(x_1,\cdots,x_n,y)\log\frac{p(x_1,\cdots,x_n,y)}{p(x_1,\cdots,x_n)\,p(y)}dx_1\cdots dx_n dy$$

其次，进行最大相关性搜索，所选变量需满足公式如下：

$$\max D(X,y),\quad D=\frac{1}{|X|}\sum_{x_i\in X}I(x_i;y) \quad (5\text{-}3)$$

通过最大相关性搜索的变量可能具有冗余性，为分析变量之间的依赖性，则使用最小依赖性剔除冗余变量。

$$\max\Phi(D,R),\quad \Phi=D-R \quad (5\text{-}4)$$

对于降低辛烷值损失的最大影响变量，采用增量搜索算法，最大化最大相关性和最小冗余性，选出最优主变量集合，其公式如下：

$$\max_{x_j\in X\text{-}X_{n\text{-}1}}\left[I(x_j;y)-\frac{1}{n-1}\sum_{x_i\in X_{n-1}}I(x_i;x_j)\right] \quad (5\text{-}5)$$

5.3.2 增量特征选择与高斯核函数支持向量回归模型（rbf-SVR）建立

【模型建立】

增量特征选择方法，是将求解目标函数最优解的一个过程。本题中，将拟合优度（R^2）、均方误差（MSE）、平均绝对误差（MAE），3 个参数作为增量特征选择的求解目标。

根据 mRMR 特征选择方法求解的 n 个主向量候选排名集（本题中降维要求 n 小于 30），使用高斯核函数支持向量回归模型，求解主要变量集合 $X_k\ (k\in[1,n])$。依次迭代求解，最大拟合优度、最小均方误差和最小平均绝对误差，其公式如下：

$$\max R,\quad R=\sum_1^k R_i,\quad i=1,2,\cdots,k \quad (5\text{-}6)$$

$$\min MSE,\quad MSE=\sum_1^k MSE_i,\quad i=1,2,\cdots,k \quad (5\text{-}7)$$

$$\min MAE,\quad MAE=\sum_1^k MAE_i,\quad i=1,2,\cdots,k \quad (5\text{-}8)$$

本题的模型使用高斯核函数-支持向量回归模型(rbf-SVR) [7]，来求解最优辛烷值损失的主要变量集合。X 作为自变量集合，辛烷值为因变量，设定样本数据为 $D=(X,Y)$，其对应的出现辛烷值损失输出记为 $f(x)$，使得其与 y 尽可能的接近，与真实的输出 y 的差别来计算损失。

假定 $f(x)$ 与 y 之间最大偏差值为 ε。w,b 是待确定的参数。模型的理想状态下中，只有当 $f(x)$ 与 y 完全相同时，其预测数量为完全正确。而当且仅当 $f(x)$ 与 y 的差的绝对值

大于ε时，才计算偏差度，此时相当于以$f(x)$为中心，构建一个宽度为2ε的预测圈，若预测数量落入此预测圈，则认为是被预测正确的。（预测圈内外的松弛程度可有所不同）。

因此，支持向量回归模型(SVR)可转化为（下式左部是正则化项）：

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}l(f(x_i)-y_i) \tag{5-9}$$

l为损失函数

$$l_\varepsilon(z)=\begin{cases}0, & if|z|\le\varepsilon\\ |z|-\varepsilon, & otherwise\end{cases} \tag{5-10}$$

因此引入了松弛因子，重写第一个式子为：

$$\begin{aligned}&\min_{w,b,\xi_i,\hat{\xi}_i}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)\\ &s.t.\ \ f(x_i)-y_i\le\varepsilon+\xi_i,\\ &\qquad y_i-f(x_i)\le\varepsilon+\hat{\xi}_i,\\ &\qquad \xi_i\ge0,\hat{\xi}_i\ge0,i-1,2,\cdots,m\end{aligned} \tag{5-11}$$

最后引入拉格朗日乘子，可得拉格朗日函数：

$$\begin{aligned}&L(w,b,\alpha,\hat{\alpha},\xi,\hat{\xi},\mu,\hat{\mu})\\ &=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)-\sum_{i=1}^{m}\mu_i\xi_i-\sum_{i=1}^{m}\hat{\mu}_i\hat{\xi}_i\\ &+\sum_{i=1}^{m}\alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)+\sum_{i=1}^{m}\hat{\alpha}_i(y_i-f(x_i)-\varepsilon-\hat{\xi}_i)\end{aligned} \tag{5-12}$$

对四个遍历求偏导，令偏导数为零，可得

$$\begin{aligned}&w=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i,\\ &0=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i),\\ &C=\alpha_i+\mu_i,\\ &C=\hat{\alpha}_i+\hat{\mu}_i.\end{aligned} \tag{5-13}$$

把上边的式子带入，即可求得SVR的对偶问题

$$\begin{aligned}&\max_{\alpha,\hat{\alpha}}\sum_{i=1}^{m}y_i(\hat{\alpha}_i-\alpha_i)-\varepsilon(\hat{\alpha}_i+\alpha_i)-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\hat{\alpha}_i-\alpha_i)(\hat{\alpha}_j-\alpha_j)x_i^Tx_j\\ &s.t.\ \sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)=0\\ &\qquad 0\le\alpha_i,\hat{\alpha}_i\le C.\end{aligned} \tag{5-14}$$

上式的过程需要满足拉格朗日乘子条件，即

$$\begin{cases} \alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)=0, \\ \hat{\alpha}_i(y_i-f(x_i)-\varepsilon-\hat{\xi}_i)=0, \\ \alpha_i\hat{\alpha}_i=0,\xi_i\hat{\xi}_i=0, \\ (C-\alpha_i)\xi_i=0,(C-\hat{\alpha}_i)\hat{\xi}_i=0. \end{cases} \tag{5-15}$$

最后，可得 SVR 的解为

$$f(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^T x+b. \tag{5-16}$$

其中 b 为

$$b=y_i+\varepsilon-\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^T x. \tag{5-17}$$

【算法流程】

表 5.1 主要变量筛选算法流程

输入：	样本集 D(X,Y), mRMR 变量列表(n)
过程：	
1	评估参数 R^2, MSE, MAE
2	For k∈ $[1,n]$ do
3	初始化 k 个变量与 y 值；
4	数据预处理标准化；
5	求解高斯核函数支持向量回归模型；
6	根据当前样本输出$\hat{y}$求解 R^2, MSE, MAE；
7	取 max R^2, min MSE, min MAE
输出：	输出评估参数最值时 k 的取值

5.3.3 模型评估与合理性验证方法

本题除了使用标准模型评估参数 R^2,MSE,MAE 进行评估意外以外，还使用了相关系数、互信息和最大信息系数（MIC）[6]、递归特征消除（RFE）[8]、随机森林（RF）[9]，对主要变量筛选进行重要性平稳度的合理性验证。

$$MAE(y,\hat{y})=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{5-18}$$

$$MAE(y,\hat{y})=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{5-19}$$

$$R^2\left(y,\hat{y}\right)=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2},\quad \text{其中}\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{5-20}$$

（1）相关系数：

$$r\left(X,Y\right)=\frac{Cov\left(X,Y\right)}{\sqrt{Var\left[X\right]Var\left[Y\right]}}$$

其中，$Cov\left(X,Y\right)$为X与Y的协方差，$Var\left[X\right]$为X的方差，$Var\left[Y\right]$为Y的方差。

（2）最大信息系数 MIC：

$$MIC\left[x;y\right]=\max_{|X||Y|<B}\frac{I\left[X;Y\right]}{\log_2\left(\min\left(|X|,|Y|\right)\right)} \tag{5-21}$$

（3）递归特征消除 RFE：

主要思想是针对那些特征含有权重的预测模型，RFE 通过递归的方式，不断减少特征集的规模来选择需要的特征。第一：给每一个特征指定一个权重，接着采用预测模型在这些原始的特征上进行训练。第二：在获取到特征的权重值后，对这些权重值取绝对值，把最小绝对值剔除掉。第三：按照这样做，不断循环递归，直至剩余的特征数量达到所需的特征数量。

（4）随机森林主要步骤 RF

①计算每个特征的重要性，并按降序排序；

②确定要剔除的比例，依据特征重要性剔除相应比例的特征，得到一个新的特征集；

③用新的特征集重复上述过程，直到剩下 m 个特征（m 为提前设定的值）；

④根据上述过程中得到的各个特征集和特征集对应的袋外误差率，选择袋外误差率最低的特征集。

5.4 主要变量筛选的模型求解与合理性验证

5.4.1 最大相关—最小冗余特征候选集求解

本题使用 Linux 系统下的 mRMR 工具[10]进行运算，将主变量候选集合 n 值设定为 30，所得到的最大特征—最小冗余分数列表如图 5.2 所示；

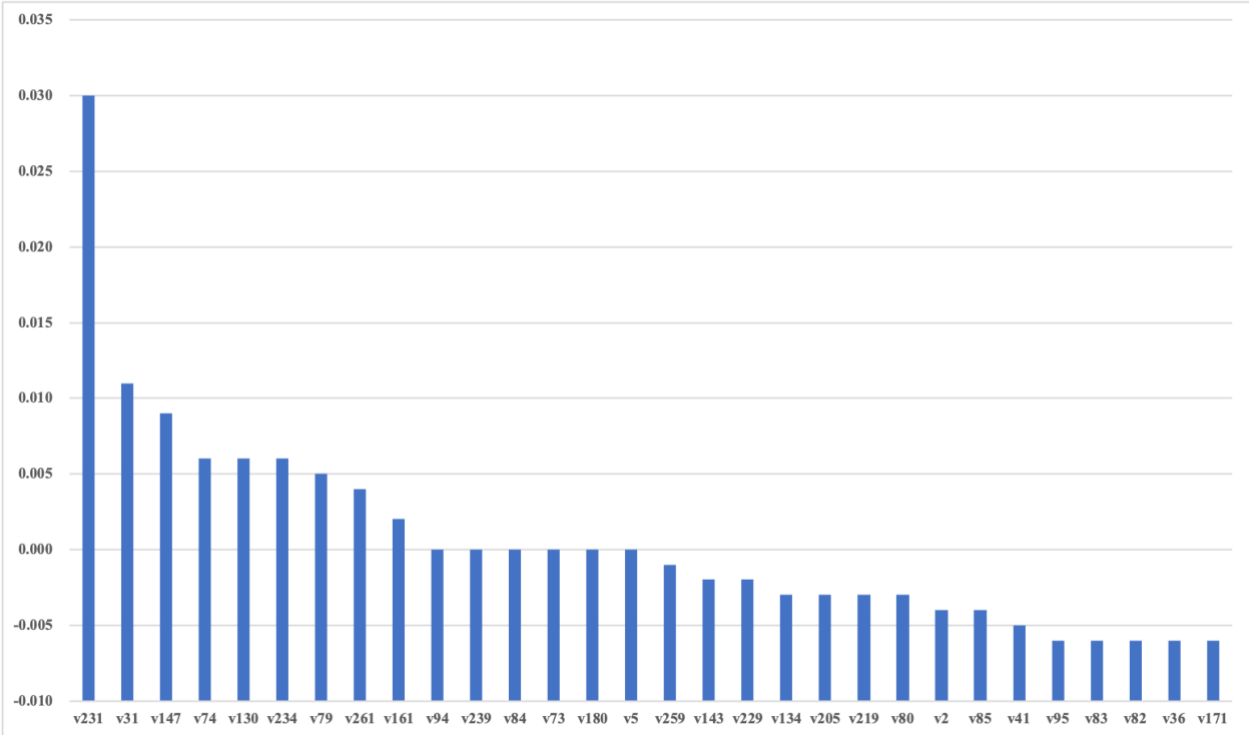


图 5.2 主要变量候选集合

表 5.2 30 个主要变量的候选列表

序号	变量编号	变量来源	变量名称
1	v231	S-ZORB.BS_AT_2401.PV	闭锁料斗烃含量
2	v31	S-ZORB.TE_5202.PV	精制汽油出装置温度
3	v147	S-ZORB.LC_3301.DACA	D123 冷凝水罐液位
4	v74	S-ZORB.PT_6002.PV	加热炉炉膛压力
5	v130	S-ZORB.TE_1104.DACA	E-101F 管程出口管温度
6	v234	S-ZORB.FT_3702.DACA	闭锁料斗 H2 过滤器出口气流量
7	v79	S-ZORB.LC_1201.PV	D104 液面
8	v261	S-ZORB.PDT_3502.DACA	ME-109 过滤器差压
9	v161	S-ZORB.LT_2901.DACA	D-109 吸附剂料位
10	v94	S-ZORB.TC_2607.PV	再生器温度
11	v239	S-ZORB.PDT_2606.DACA	R-102 底滑阀差压
12	v84	S-ZORB.FC_1203.PV	D121 去稳定塔流量
13	v73	S-ZORB.TC_1606.PV	反应器入口温度
14	v180	S-ZORB.TE_5004.DACA	稳定塔顶出口温度
15	v5	原料性质	芳烃,v%
16	v259	S-ZORB.PDT_2409.DACA	ME-115 过滤器压差
17	v143	S-ZORB.LT_3101.DACA	D-124 液位
18	v229	S-ZORB.BS_AT_2402.PV	闭锁料斗氧含量
19	v134	S-ZORB.LI_9102.DACA	D-204 液位
20	v205	S-ZORB.PC_3001.DACA	D-113 压力
21	v219	S-ZORB.FT_2803.DACA	紧急氢气去 D-102 流量
22	v80	S-ZORB.FC_1201.PV	D104 去稳定塔流量
23	v2	原料性质	辛烷值 RON

24	v85	S-ZORB.FC_1202.PV	D121 顶去放火炬流量
25	v41	S-ZORB.PT_9001.PV	燃料气进装置压力
26	v95	S-ZORB.AI_2903.PV	再生烟气氧含量
27	v83	S-ZORB.LC_1202.PV	D121 液面
28	v82	S-ZORB.TE_1203.PV	D121 温度
29	v36	S-ZORB.FT_1501.PV	新氢进装置流量
30	v171	S-ZORB.FT_2433.DACA	D-106 压力仪表管嘴反吹气流量

从最大相关—最小冗余特征选择方法的结果中可以看出，“闭锁料斗烃含量”对于降低辛烷值的影响最大；其次，除了操作变量之外，原料性质中“芳烃”和“辛烷值”也对辛烷值的降低也具有一定的影响。

5.4.2 增量特征选择与高斯核函数支持向量回归模型的求解

基于 mRMR 计算得出的主变量候选列表，我们联用增量特征选择，以及适用于非线性变量的高斯核函数支持向量回归模型，以最高的拟合优度（max R^2）、最低均方误差（min MSE）、最低平均绝对误差（min MAE），进一步得出最终的主要变量筛选结果。

本题主要使用 Python 编写的“test2-1.py”脚本，进行 30 次迭代运算，分别计算根据 mRMR 变量列表，1 到 30 个变量对于预测辛烷值损失的最佳精确度（即最高拟合优度、最低均方误差和最低平均绝对误差）。

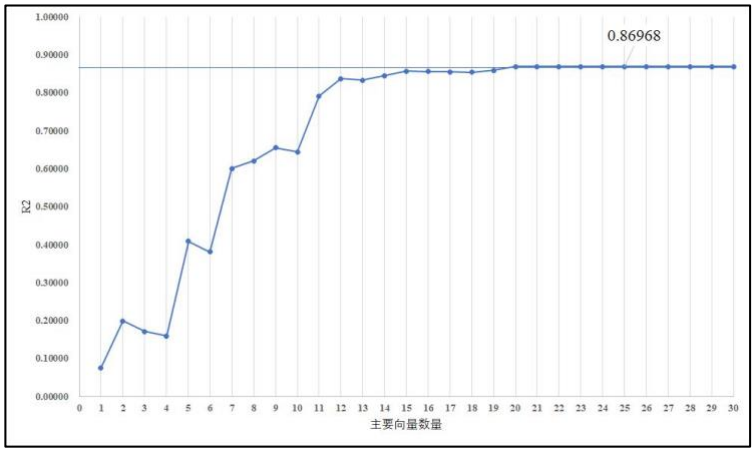


图 5.3 拟合优度

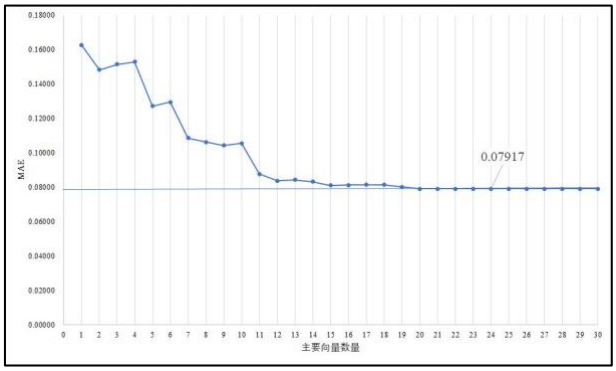


图 5.4 MSE

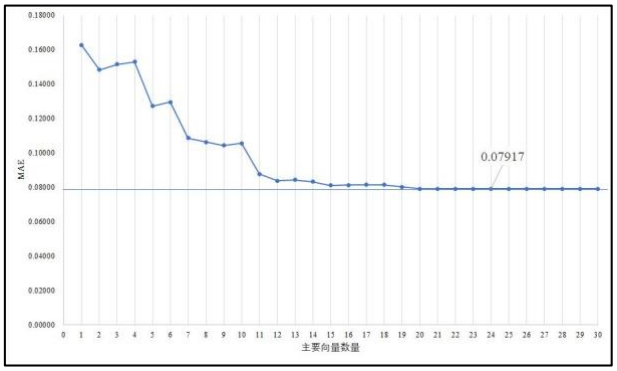


图 5.5 MAE

【主要变量筛选结果】

从图中可以看出，根据 mRMR 主变量候选列表，**前 24 个**变量的候选集合的拟合优度最高，且均方误差和平均绝对误差最低，分别为 **86.97%、0.0069、0.0792**，能够达到很好

的预测效果。因此最终的主要变量筛选结果如下表 5.3 所示。

表 5.3 主要变量筛选结果

序号	变量编号	变量代号	变量名称
1	v231	S-ZORB.BS_AT_2401.PV	闭锁料斗烃含量
2	v31	S-ZORB.TE_5202.PV	精制汽油出装置温度
3	v147	S-ZORB.LC_3301.DACA	D123 冷凝水罐液位
4	v74	S-ZORB.PT_6002.PV	加热炉炉膛压力
5	v130	S-ZORB.TE_1104.DACA	E-101F 管程出口管温度
6	v234	S-ZORB.FT_3702.DACA	闭锁料斗 H2 过滤器出口气流量
7	v79	S-ZORB.LC_1201.PV	D104 液面
8	v261	S-ZORB.PDT_3502.DACA	ME-109 过滤器差压
9	v161	S-ZORB.LT_2901.DACA	D-109 吸附剂料位
10	v94	S-ZORB.TC_2607.PV	再生器温度
11	v239	S-ZORB.PDT_2606.DACA	R-102 底滑阀差压
12	v84	S-ZORB.FC_1203.PV	D121 去稳定塔流量
13	v73	S-ZORB.TC_1606.PV	反应器入口温度
14	v180	S-ZORB.TE_5004.DACA	稳定塔顶出口温度
15	v5	原料性质	芳烃,v%
16	v259	S-ZORB.PDT_2409.DACA	ME-115 过滤器压差
17	v143	S-ZORB.LT_3101.DACA	D-124 液位
18	v229	S-ZORB.BS_AT_2402.PV	闭锁料斗氧含量
19	v134	S-ZORB.LI_9102.DACA	D-204 液位
20	v205	S-ZORB.PC_3001.DACA	D-113 压力
21	v219	S-ZORB.FT_2803.DACA	紧急氢气去 D-102 流量
22	v80	S-ZORB.FC_1201.PV	D104 去稳定塔流量
23	v2	原料性质	辛烷值 RON
24	v85	S-ZORB.FC_1202.PV	D121 顶去放火炬流量

从筛选结果可以看出，各个变量对辛烷值损失降低的影响，主要体现在以下几个方面：

（1）所有样本的 24 个主变量均在变量规范的范围之内，并且包含了题目中需纳入的原料“辛烷值”这一建模变量；

（2）变量筛选过程中，剔除了样本中一些本身存在问题（变量值在范围之外）的变量“原料缓冲罐液位”、“ 再生烟气氧含量”、“ S-ZORB.FT_1204.TOTAL”；

5.4.3 合理性验证结果与分析

为验证变量筛选的合理性，我们主要使用了相关系数、互信息和最大信息数、随机森林、递归特征消除等方法，代码文件见附件“test2-2.py”。由于数据中操作变量的具有高度非线性和相互强耦联的关系，并且数据中存在跨度大，不在标准规定的范围之内，所以用多种检验方式来体现主要变量在不同规则下的重要平稳度。

表 5.4 多方法检验值比对

变量编号	相关系数	RF	**MIC**	**RFE**
v2	0.00539	0.09621	0.25817	0.77778
v5	0.03229	0.04902	0.49234	0.84722
v30	0.00551	0.001	0.54899	0.46944
v73	0.01115	0.00834	0.48959	0.625

v74	0.22955	0.05569	0.6203	0.98611
v79	0.00696	0.02127	0.43846	0.56667
v80	0.00501	0.00694	0.71052	0.52778
v84	0.03753	0.02496	0.60456	0.73333
v85	0.0022	0.00391	0.43317	0.13611
v94	0.00927	0.0303	0.51089	0.62778
v130	0.00306	0.02375	0.52981	0.63333
v134	0.00181	0.19886	0.51893	0.29167
v143	0.00303	0.05946	0.76983	0.65278
v147	0.04014	0.07131	0.57458	0.56111
v161	0.03126	0.15974	0.64809	0.225
v180	0.01771	0.04694	0.59087	0.49167
v204	0.001	0	0.63386	0.52222
v219	0.15023	0	0.10543	0.83056
v229	0.0818	0.26292	0.73427	0.36944
v231	0.00009	0	0.57811	0.42778
v234	0.14318	0	0.61632	0.04167
v239	0.05873	0	0.61993	0.18333
v259	0.02933	0.00377	0.65272	0.43611
v261	0.0184	0.44987	0.59215	0.53056

从图中可以看出，24 个主要变量在四种衡量变量重要性的异同。可以看出，主要变量在多种方法下的评价得分规律：

（1）互信息—最大信息数 MIC、递归特征消除评分 REF 评分较为相近，体现了主要变量在不同规则下的平稳度；

（2）MIC 和 REF 能够显示出主要变量的重要性和显著度；

（3）各个变量与辛烷值损失的线性相关性比较低，所以预测辛烷值损失更适合使用非线性的回归模型进行预测；

（4）主要变量在随机森林方法下的评分过于离散，不适用于少量样本；

主要变量具有平稳度、显著度，验证了主要变量筛选的合理性。

5.5 小结与讨论

本题筛选主要变量时，存在两个主要的难题：

（1）操作变量之间具有高度非线性和相互强耦联性，一般的线性模型无法直接衡量变量的相关性和重要性；

（2）样本数据量较少，不适用于大样本模型的应用；

（3）在使用非线性模型评价变量的主要性时，可能存在多种局部最优的组合，所以需要结合实际情况来具体分析。

六、辛烷值损失预测模型的建立与求解

6.1 符号约定

符号	含义
D	数据样本集合
$Gini(D)$	D 的基尼指数
L	损失函数
$H(X)$	衡量节点不纯度的函数

6.2 问题分析

由于催化裂化汽油精制过程是连续的，所以实际情况中，可以通过操作过程降低辛烷值损失。同时，原料性质、待生吸附剂性质、再生吸附剂性质也可能对辛烷值有所影响。因此，通过筛选影响辛烷值降低的主要变量，能够进一步推进预测辛烷值模型的建立，为后续的汽油实际精制奠定基础。

本题主要通过数据挖掘技术建立辛烷值损失预测模型，即通过样本数据，求解最佳回归模型。由于数据本身的变量之间具有高度非线性和相互强耦联的关系，因此，本题选取相应的回归模型进行求解。我们主要联用第二问的高斯核函数支持向量回归模型、决策树回归模型、随机森林回归模型、梯度上升回归树模型，比较评估参数选取最佳预测模型，同时，对预测的辛烷值损失、样本的辛烷值损失进行可视化处理与分析。

6.3 数据分析与预处理

为建立辛烷值损失模型，本题使用第二问中筛选的 24 个主要变量，进行回归预测。由于数据中的变量量纲有所不同，我们使用 L1 正则化方式对数据进行预处理，同时也缓解了模型训练过程中的过拟合问题。

6.4 辛烷值损失模型的建立

在第二问的基础上，使用常用的数据挖掘技术建立预测模型。本题使用四种回归模型进行预测比较，包括支持向量回归模型、随机森林回归模型、梯度提升回归模型、决策树回归。主要从R^2、MSE、MAE 三个参数，比对选取最佳辛烷值损失模型。

（1）随机森林回归模型

RFR 是由多个二叉决策树（即 CART）打包组合而成的[11]，在训练二叉决策树模型的时候需要考虑怎样选择切分变量(特征）、切分点以及怎样衡量一个切分变量、切分点的好坏。针对于切分变量和切分点的选择，一般是通过遍历每个特征和每个特征的所有取值，从中找出最好的切分变量和切分点。以切分后节点的不纯度来衡量，即各个子节点不纯度的加权和，其计算公式如下：

$$G\left(x_i, v_{ij}\right)=\frac{n_{left}}{N_S} H\left(X_{left}\right)+\frac{n_{right}}{N_S} H\left(X_{right}\right) \qquad (6\text{-}1)$$

其中，x_i 为某一个切分变量，v_{ij} 为切分变量的一个切分值，n_{left}，n_{right}，N_s 分别为切分后左子节点的训练样本个数、右子节点的训练样本个数以及当前节点所有训练样本个数，X_{left}，X_{right} 分为左右子节点的训练样本集合，$H(X)$为衡量节点不纯度的函数(impurity function/criterion)，分类和回归任务一般采用不同的不纯度函数。

本文在回归采用的不纯度函数为平方平均误差（MSE）和绝对平均误差（MAE），公式如下：

平方平均误差（MSE）函数公式：
$$H\left(X_m\right)=\frac{1}{N_m}\sum_{i\in N_m}\left(y-\overline{y_m}\right)^2 \tag{6-2}$$

绝对平均误差（MAE）函数公式：
$$H\left(X_m\right)=\frac{1}{N_m}\sum_{i\in N_m}\left|y-\overline{y_m}\right|^2 \tag{6-3}$$

（2）决策树回归模型

CART 决策树使用“基尼指数”来选择划分属性[11]。数据集为$D=\left(X,Y\right)$，集合 D 中第 n 类样本所占的比例为 $p_i\left(i=1,2,\cdots,\left|n\right|\right)$，则 D 的纯度可用基尼值来度量：

$$Gini\left(D\right)=\sum_{i=1}^{\left|n\right|}\sum_{i'\neq i}p_i\,p_{i'}=1-\sum_{i=1}^{\left|n\right|}p_i^2 \tag{6-4}$$

直观来说，$Gini\left(D\right)$反映了从数据集 D 中随机抽取两个样本，其类别标记不一致的概率。因此，$Gini\left(D\right)$越小，则数据集 D 的纯度越高。

假定离散属性 x 有 V 个可能的取值$\left\{x^1,x^2,\cdots,x^v\right\}$,若使用 x 来对样本集 D 进行划分，则会产生 V 个分支结点，其中第 v 个分支结点包含了 D 中所有在属性 x 上取值为 x^v 的样本，记为D^v。再考虑到不同的分支结点所包含的样本数不同，给分支结点赋予权重$\left|D^v\right|/\left|D\right|$，即样本越多的分支结点的影响越大。于是属性 x 的基尼指数定义为：

$$Gini\_index\left(D,x\right)=\sum_{v=1}^{V}\frac{\left|D^v\right|}{\left|D\right|}Gini\left(D^v\right) \tag{6-5}$$

于是，我们在候选属性集合 X 中，选择那个使得划分后基尼指数最小的属性作为最优划分属性，即：

$$x_*=\arg\min_{x\in X}Gini\_index\left(D,x\right) \tag{6-6}$$

（3）梯度提升回归树模型

设定样本数据为$D=\left(X,Y\right)$，其对应的出现辛烷值损失输出记为 $f\left(x\right)$，使得其与 y 尽可能的接近，与真实的输出 y 的差别来计算损失[12]。

损失函数 L 为：

$$L\left(y,f\left(x\right)\right)=\left(y-f\left(x\right)\right)^2 \tag{6-7}$$

1) 初始化弱学习器：

$$f_0\left(x\right)=\arg\min_c\sum_{i=1}^{n}L\left(y_i,c\right) \tag{6-8}$$

2) 对于迭代轮数$m=1,2,\cdots,M$ 有：

对样本计算负梯度：

$$r_{mi}=-\left[\frac{\partial L\left(y_i,f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right]f_{m-1}\left(x\right) \tag{6-9}$$

利用$\left(x_i,r_{mi}\right),\left(i=1,2,\cdots,n\right)$ 拟合一棵 CART 回归树，得到第 m 棵回归树，其对应的叶

子结点区域为$R_{mk}, k=1,2,\cdots,K$，其中 K 为回归树 m 的叶子结点个数。

对叶子区域计算出最佳拟合值：

$$c_{mk}=\arg\min_{c}\sum_{x_i\in R_{mk}}L\left(y_i, f_{m-1}\left(x_i\right)+c\right) \tag{6-10}$$

更新强学习器：

$$f_m\left(x\right)=f_{m-1}\left(x\right)+\sum_{k=1}^{K}c_{mk}I\left(x\in R_{mk}\right) \tag{6-11}$$

3) 得到强学习器 $f\left(x\right)$的表达式：

$$f\left(x\right)=f_M\left(x\right)=f_0\left(x\right)+\sum_{m=1}^{M}\sum_{k=1}^{K}c_{mk}I\left(x\in R_{mk}\right) \tag{6-12}$$

本题的建模及求解流程如图 6.1 所示。

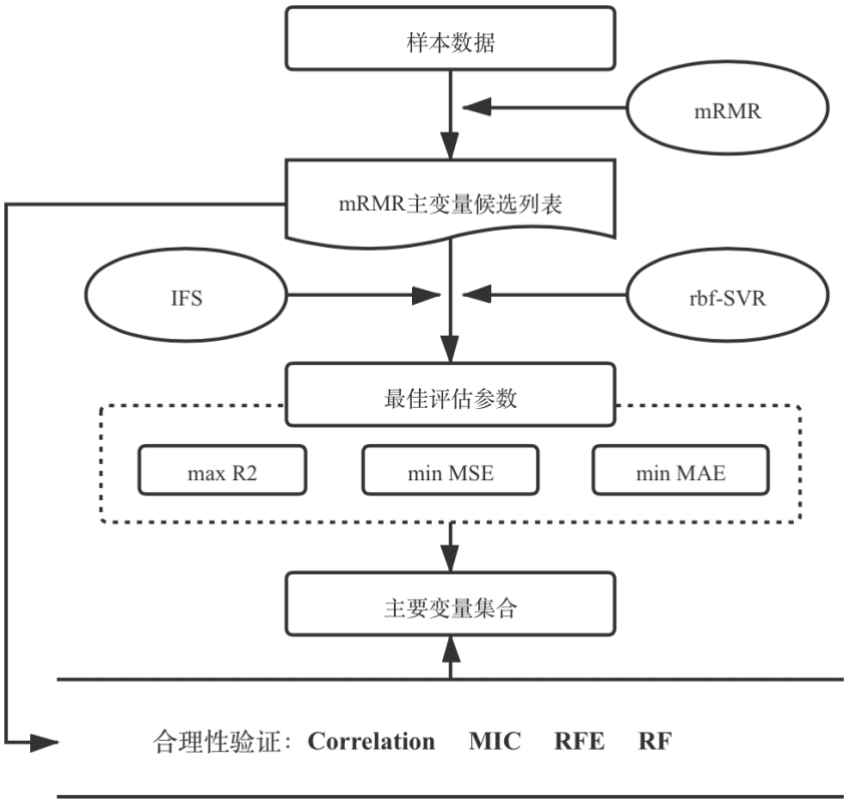


图 6.1 辛烷值损失预测建模流程

6.5 辛烷值损失模型的求解

本题主要使用常用的数据挖掘技术建立预测模型，使用四种回归模型进行预测比较，包括支持向量回归模型、随机森林回归模型、梯度提神回归模型、决策树回归。主要从R^2、MSE、MAE 三个参数，比对选取最佳辛烷值损失模型。

【模型比较】

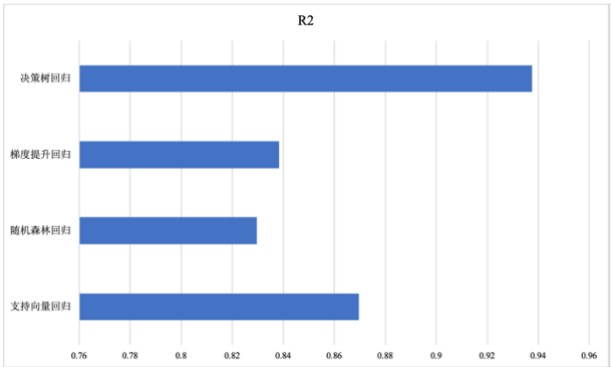


图 6.2 拟合优度比较

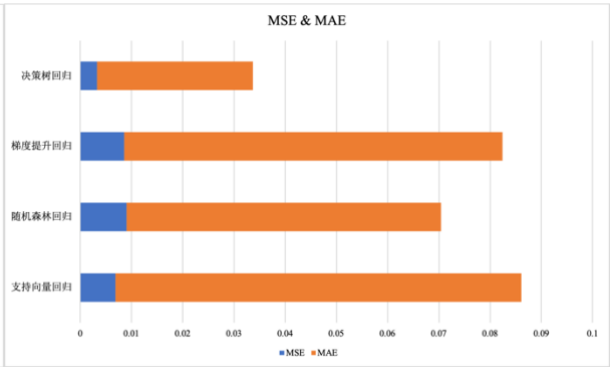


图 6.3 MSE 和 MAE 柱状堆积图比较

从图中可以看出，**决策树回归模型的拟合优度最高为 93.75%**，且 MSE、MAE 最低分别为 0.0033、0.03046，因此选取决策树回归模型，作为辛烷值损失预测模型。

【模型预测】

本题使用 python 编程，可视化预测结果。我们将各个样本预测的值相连，更能直观得看出预测值与样本实际值的预测差距。四种回归模型的辛烷值损失预测结果如图所示：

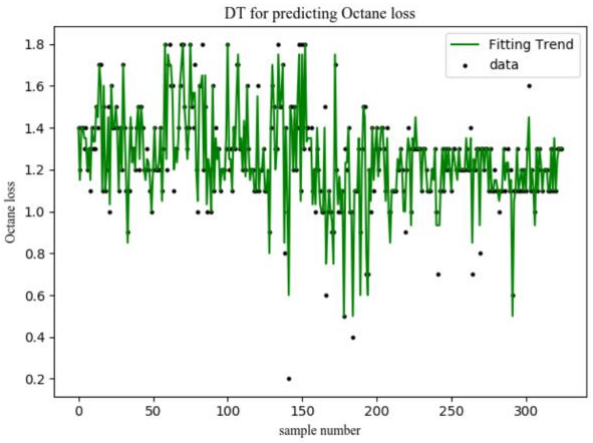


图 6.4 决策树回归预测

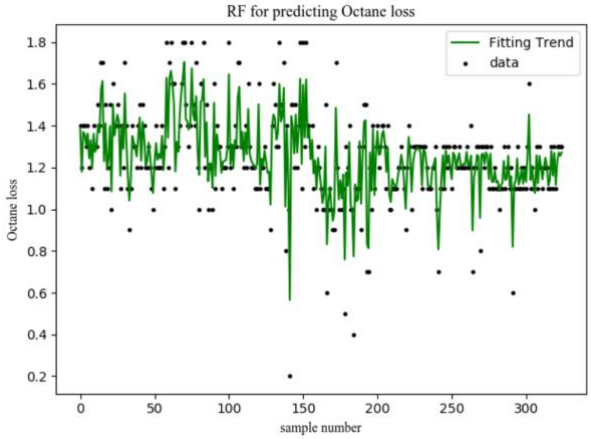


图 6.5 随机森林回归预测

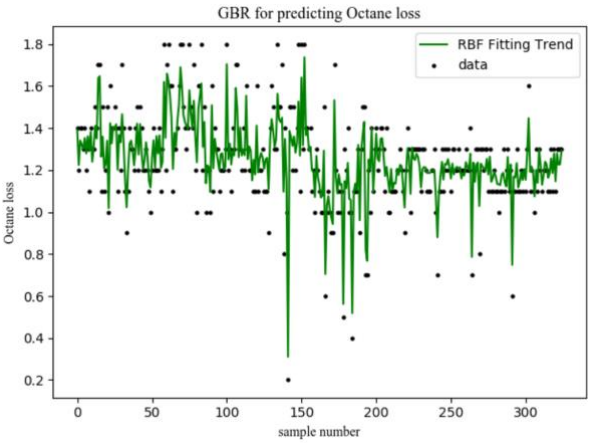


图 6.6 梯度提升回归树预测

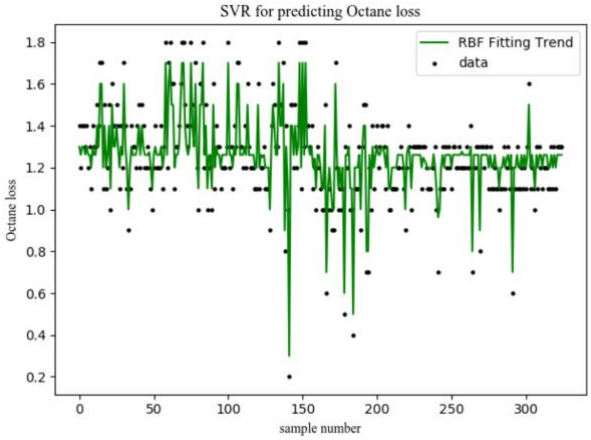


图 6.7 支持向量回归预测

四种模型的辛烷值损失预测数据——见**附录“6-1 回归预测结果”**。

本题使用网格搜索迭代选取决策树回归的最优参数。其中，使用 L1 范式防止过拟合，参数设置 min_samples_split=2, min_samples_leaf=2。从回归预测中可以看出，决策树的预测结果优度最高。

6.6 小结与讨论

本题辛烷值损失值不够精确，并且操作变量存在些许问题，如部分数据不在规定范围之内等。本题总结点可归纳如下：

（1）在进行数据挖掘的时候，需要更多更为精确的数据进行建模，可为相关企业在汽油精制过程中，降低辛烷值损失奠定良好的基础，提供更为确切的实行方案。

（2）决策树更适用于预测样本数据的辛烷值损失，可能的潜在原因是在于辛烷值值损失精确到小数点后一位，在根据变量决策时，更容易进行回归建模。

七、操作方案建模与优化

7.1 问题分析

现有技术在对催化裂化汽油进行脱硫和降烯烃过程中，普遍降低了汽油辛烷值，造成了辛烷值（RON）的损失。由此可知，降低产品硫含量和减少辛烷损失为存在冲突。本研究主要目的在于保证产品硫含量不大于 5μg/g 的前提下，分析辛烷值（RON）损失降幅大于 30%的样本对应的主要变量优化后的操作条件。考虑到优化过程中原料、待生吸附剂、再生吸附剂的性质保持不变，因此以它们在样本中的数据为准。

基于合作博弈论[13]，建立目标优化准则。首先，我们对筛选样本数据进行分析，研究主要变量分布；其次，建立多目标优化方案，进行约束——主要变量取值范围约束、产品硫含量约束、辛烷损失降幅约束。

建立“合作博弈——PareTo 强度进化算法”[14]。首先，输入样本数据的操作变量，约束其上下限，根据“合作博弈”方法、“PareTo 强度进化算法”，进行多次迭代求最优解。获取多次迭代的最优解对应的操作变量值，使用多种密度函数求得均值以及置信区间。最后，通过验证样本契合度、以及 95%的置信区间，说明模型的合理性。

本题使用的主要要变量对应信息表如下：

表 7.1 主要变量筛选结果

序号	变量编号	变量代号	变量名称
1	v231	S-ZORB.BS_AT_2401.PV	闭锁料斗烃含量
2	v31	S-ZORB.TE_5202.PV	精制汽油出装置温度
3	v147	S-ZORB.LC_3301.DACA	D123 冷凝水罐液位
4	v74	S-ZORB.PT_6002.PV	加热炉炉膛压力
5	v130	S-ZORB.TE_1104.DACA	E-101F 管程出口管温度
6	v234	S-ZORB.FT_3702.DACA	闭锁料斗 H2 过滤器出口气流量
7	v79	S-ZORB.LC_1201.PV	D104 液面
8	v261	S-ZORB.PDT_3502.DACA	ME-109 过滤器差压
9	v161	S-ZORB.LT_2901.DACA	D-109 吸附剂料位
10	v94	S-ZORB.TC_2607.PV	再生器温度
11	v239	S-ZORB.PDT_2606.DACA	R-102 底滑阀差压
12	v84	S-ZORB.FC_1203.PV	D121 去稳定塔流量
13	v73	S-ZORB.TC_1606.PV	反应器入口温度
14	v180	S-ZORB.TE_5004.DACA	稳定塔顶出口温度
15	v5	原料性质	芳烃,v%
16	v259	S-ZORB.PDT_2409.DACA	ME-115 过滤器压差
17	v143	S-ZORB.LT_3101.DACA	D-124 液位
18	v229	S-ZORB.BS_AT_2402.PV	闭锁料斗氧含量
19	v134	S-ZORB.LI_9102.DACA	D-204 液位
20	v205	S-ZORB.PC_3001.DACA	D-113 压力
21	v219	S-ZORB.FT_2803.DACA	紧急氢气去 D-102 流量
22	v80	S-ZORB.FC_1201.PV	D104 去稳定塔流量
23	v2	原料性质	辛烷值 RON
24	v85	S-ZORB.FC_1202.PV	D121 顶去放火炬流量

7.2 筛选样本数据，分析主要变量分布

基于产品中硫含量不大于 5 μg/g 的约束条件，对 325 条样本数据进行筛选，研究产品硫含量较低时，其主要变量的特征。经过筛选，有 57 条数据超过约束条件，268 条数据低于 5 μg/g。对两类数据的 24 个主要变量的分布进行对比，可得：

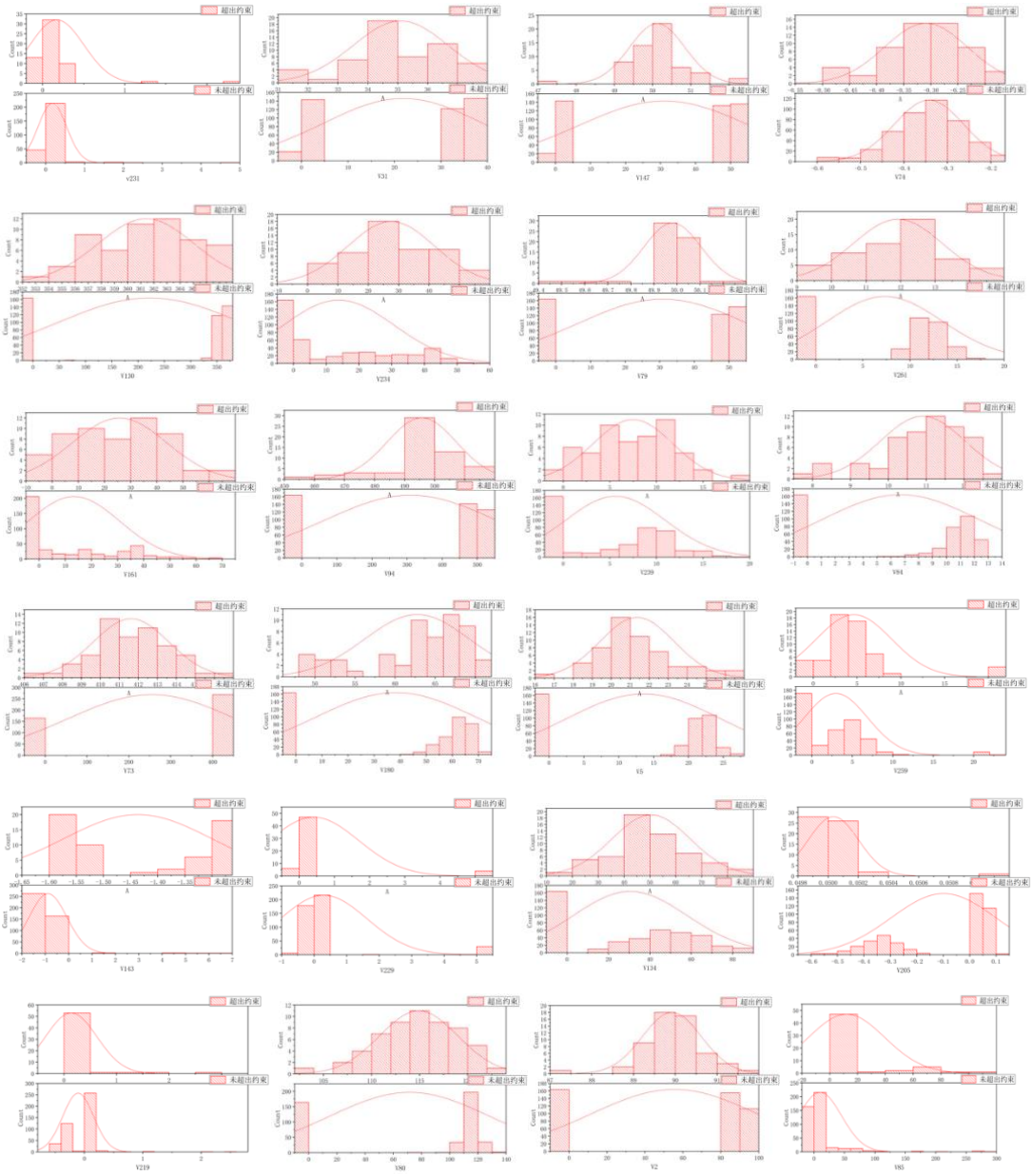


图 7.1 样本数据主要变量分布（大图见附件“7-1”）

从 24 个主要变量的分布图未超出约束和超出约束两类数据，总体较为集中，因此不能通过单一调整某一参数：从平均值，从正负值，更服从正态分布，从相关性正负验证。从变量分布来看，可以总结以下几点规律：

（1）部分变量符合标准的正态分布，对其进行操作调整达到最优解的过程中，符合条件的样本量占比可能更高。

（2）部分变量分布存在极端值，说明对其进行操作优化时，可能很快达到最优解；

（3）符合标准正态分布的变量是重要调整的对象，分布存在极端值的变量应次要调整。

7.3 基于合作博弈论，建立目标优化准则

催化裂化汽油精制过程中，在产品硫含量和辛烷值损失降幅的边界约束下，通过调整主要操作变量，使目标均尽可能最优。由于主要操作变量并非相互独立，而是相互联系、相互耦合的，需要应用多目标优化方法进行综合求解。传统方法是将多个优化目标使用一定的转换方法变换为单个目标，该方法应用简单，但原问题方法需用明确的数学表达式来描述，并对不同量纲的多个目标进行强行耦合。

而对于无法建立明确数学表达式的优化目标，需要改进传统多目标优化方法来解决。本题结合合作博弈论的内容，建立目标优化准则，对多个目标进行分解。博弈问题的构成涉及博弈参与者、各方收益、各方可支配的策略集三个主要内容。与多目标优化问题中的相关要素相对应，如图所示。

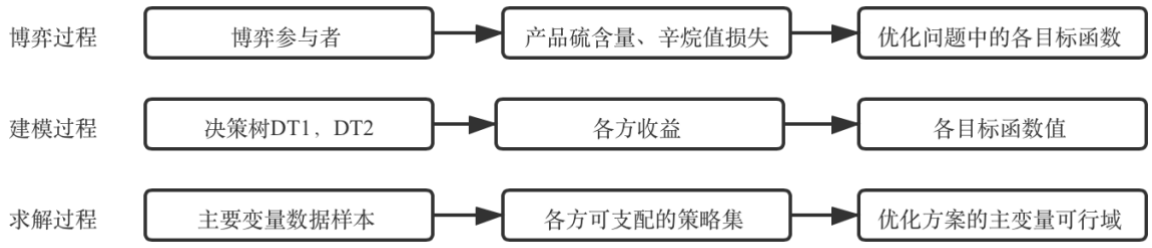


图 7.2 操作方案的合作博弈论

合作博弈论中，博弈参与者彼此间信息互通，同时决策，各方期望的目标以避免整体损失为前提，强制执行具有约束力的契约。合作博弈各方相互协商，进行合作，彼此妥协与退让，达成共识，既使得自身目标达成，又使得集体条件满足。类似地，多目标优化过程中，每个目标函数都是博弈中的一方，各目标之间进行合作协商，制定合约，并按照约定强制执行决策。结合合作博弈理论和优化目标，建立以下策略。

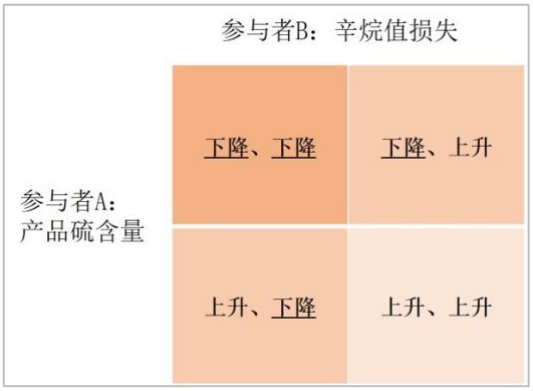


图 7.3 合作博弈策略图

在求解最优目标时，优先实现参与者均满意的情况，即产品硫含量降低时，辛烷损失也随之降低；其次，实现参与者 A（产品硫含量）和参与者 B（辛烷值损失）两者之一的目标，即一升一降；最后，避免出现参与者双方均不满意的情况，此时优化失败。

7.4 基于优化策略，建立多目标优化模型

7.4.1 优化目标

减少产品硫含量和辛烷值损失是催化裂化汽油精制的两个核心目标。脱硫和降烯烃过程中，可以有效利用重油资源，满足对汽油质量要求。辛烷值（以 RON 表示）是反映汽油燃烧性能的最重要指标，辛烷值每降低 1 个单位，相当于损失约 150 元/吨。

在满足产品硫含量不大于 5 μg/g，辛烷损失降幅大于 30%的前提下，同时调整 22 个主要变量（优化过程中原料、吸附剂的性质保持不变），使 325 个样本中尽可能多的数据满足前提，优化目标如下。

$$\min\left(\mathrm{M}-325\right) \tag{7-1}$$

其中，M 为满足产品硫含量和辛烷损失降幅的数据条数。

7.4.2 约束分析

（1）主要变量取值范围约束

由于优化过程中原料、吸附剂的性质保持不变，因此对主要变量中的操作变量进行约束。根据附件 4 中不同操作变量的取值范围，确定主要变量的上下约束，即主要变量增幅范围的上下限，约束如下式所示。

$$\min\left(X_{i,j}\right) \le X_{i,j} \le \max\left(X_{i,j}\right), \forall i \in E, \forall j \in N \tag{7-2}$$

其中，$X_{i,j}$为主要操作变量；E 为主要变量所在列号的集合，E={1,2,3,…,22}；N 为每个主要变量所在行号的集合，N={1,2,3,…,325}。

（2）产品硫含量约束

在催化裂化汽油精制过程中，需要保证产品硫含量不大于 5 μg/g，约束如下所示。

$$5 - DT_2\left(X_{i,j}\right) \le 0, \forall i \in E, \forall j \in N \tag{7-3}$$

其中，$DT_2(X_{i,j})$为决策树对主要变量的预测值；$X_{i,j}$、E、N 的意义同上。

（3）辛烷损失降幅约束

在脱硫和降烯烃过程中，会造成辛烷损失，因此降低辛烷损失降幅也变的很重要，需使辛烷值（RON）损失降幅大于 30%，约束如下所示。

$$70\%Y_j^R - DT_1\left(X_{i,j}\right) \ge 0, \forall j \in N \tag{7-4}$$

其中，Y_j^R为辛烷损失值；$DT_1(X_{i,j})$为决策树对主要变量的预测值；$X_{i,j}$、E、N 的意义同上。

7.4.3 模型建立

基于合作博弈论理念，结合以上分析，在满足产品硫含量不大于 5 μg/g，辛烷损失降幅大于 30%的前提下，使 325 个样本中尽可能多的数据满足前提，建立操作变量多目标优化模型如下：

$$\min\left(M - 325\right)$$

$$s.t.\begin{cases} \min\left(X_{i,j}\right) \le X_{i,j} \le \max\left(X_{i,j}\right), \forall i \in E, \forall j \in N \\ 5 - DT_2\left(X_{i,j}\right) \le 0, \forall i \in E, \forall j \in N \\ 70\%Y_j^R - DT_1\left(X_{i,j}\right) \ge 0, \forall j \in N \end{cases} \tag{7-5}$$

7.5 操作方案模型的求解

本题使用“**合作博弈——PareTo 强度进化算法**”[14]，求解主要变量操作的优化方案。使用 python 编程求解，见附件“test4.py”,“dealdata.py”，“ dealdata2.py”。使得尽可能多的产品样本在调整操作变量后，硫含量不大于 5μg/g，且辛烷值（RON）损失降幅大于

30%。

在求解多目标最优解的过程中，我们使用 NSGA-II 算法进行迭代求解，主要定义如下：

（1）基于本题多目标优化问题，设X ∈ D，如果¬∃X ∈ D，使得满足条件：对于$f(x)$的任意目标子函数$f'(x)$都有$f'(x) \leq f(x)$，同时至少存在一个子目标函数$f(x)$使得$'(x) \leq f(x)$，则 X 为 PareTo 强化最优解。

（2）对于解的支配关系，设X_1, X_2为解空间的解，如果对所有目标，X_1都比X_2要好，则X_1强支配X_2。如果至少存在一个目标，X_2比X_1好，X_1总体上比较好，则X_1为弱支配。

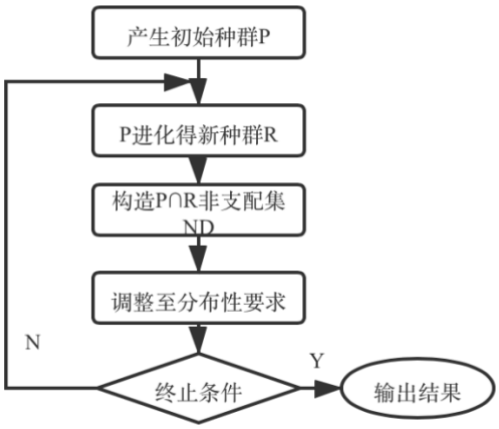


图 7.4 PareTo 强度进化算法流程

本题求解主要变量操作的优化方案中，24 个变量中 22 个为操作变量，并且具有规定的范围。因此，需尽可能满足所有样本达到多目标最优，改进了 PareTo 强度进化算法，采用多变量合作博弈，进行模型求解。

“合作博弈——PareTo 强度进化算法”流程如下图所示。首先输入样本数据的操作变量，约束其上下限，根据“合作博弈”方法、“PareTo 强度进化算法”，进行多次迭代求最优解。获取多次迭代的最优解对应的操作变量值，使用多种密度函数求得均值以及置信区间。最后，通过验证样本契合度、以及 95%的置信区间，说明模型的合理性。

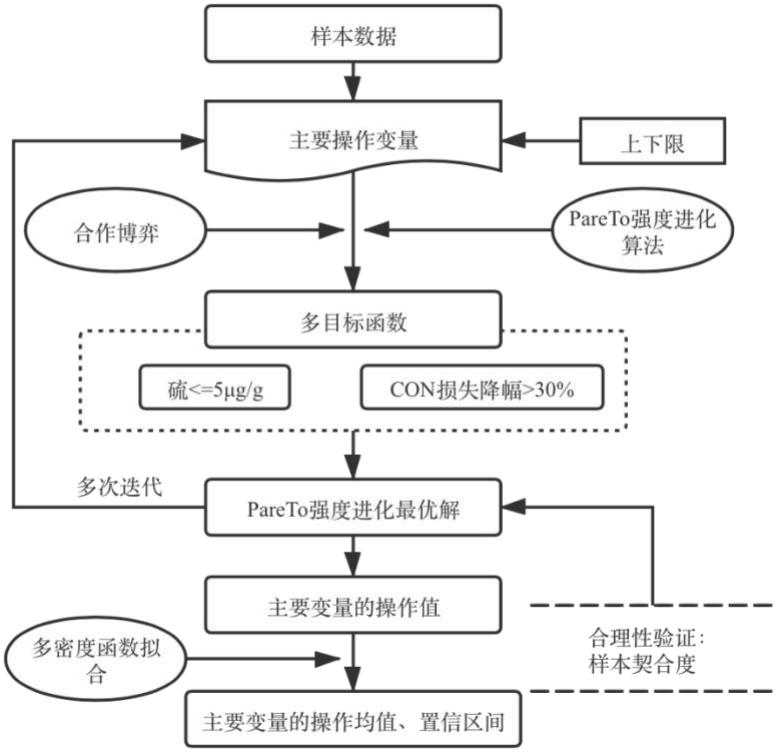
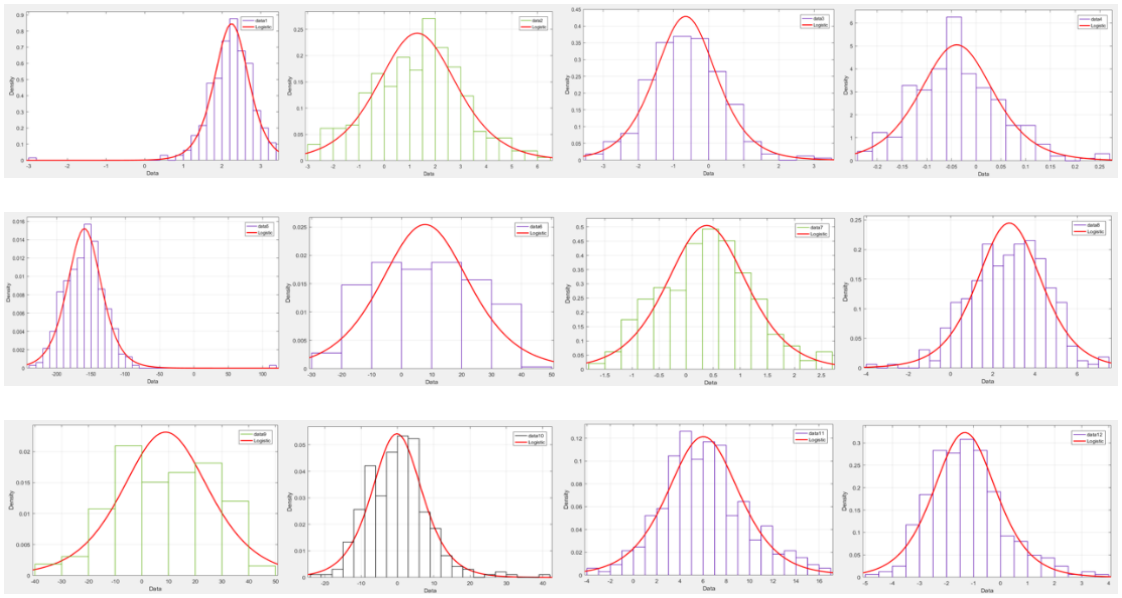


图 7.5 “合作博弈——ParetTo 强度进化算法”流程

7.5.1 主要变量分布密度函数求解

本题使用 Matlab 工具箱的密度函数拟合工具，主要使用 Logistic 函数对 PreTo 强化进度最优解对应的主要变量操作值进行拟合，拟合结果如图所示，从直方图和曲线可以看出拟合的契合度。分布密度函数的求解参数，见附件 7-2“分布密度函数的求解参数表”。



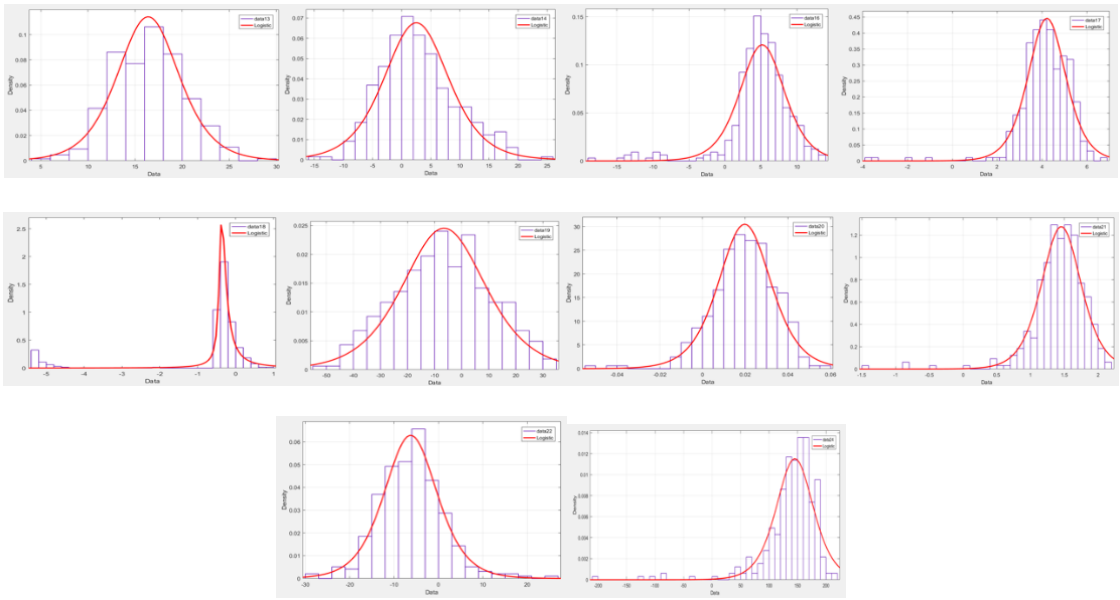


图 7.6 主要变量操作值的拟合图（大图见附件“7-3”）

经过多次迭代求解，22 个主要操作变量的分布密度拟合图，主要服从正态分布、偏态分布的情况，因此，可以从多次迭代求解的**增减幅度均值、置信区间**给出整体样本的操作优化方案。

7.5.2 主要变量操作方案优化

根据分布密度函数求解的操作变量的增减幅度均值和置信区间如下表。

表 7.2 主要变量操作方案优化

变量代号	变量名称	增减度均值	置信区间	
S-ZORB.BS_AT_2401.PV	闭锁料斗烃含量	+2.2461	1.1631	3.329
S-ZORB.TE_5202.PV	精制汽油出装置温度	+1.2913	-2.4796	5.0623
S-ZORB.LC_3301.DACA	D123 冷凝水罐液位	-0.6542	-2.7889	1.4806
S-ZORB.PT_6002.PV	加热炉炉膛压力	-0.0394	-0.2207	0.1419
S-ZORB.TE_1104.DACA	E-101F 管程出口管温度	-159.7041	-219.9004	-99.5079
S-ZORB.FT_3702.DACA	闭锁料斗 H2 过滤器出口气流量	+7.8271	-28.174	43.8282
S-ZORB.LC_1201.PV	D104 液面	+0.3798	-1.4338	2.1934
S-ZORB.PDT_3502.DACA	ME-109 过滤器差压	+2.7726	-0.9683	6.5136
S-ZORB.LT_2901.DACA	D-109 吸附剂料位	+8.8912	-30.6927	48.4752
S-ZORB.TC_2607.PV	再生器温度	-0.1347	-17.0387	16.7693
S-ZORB.PDT_2606.DACA	R-102 底滑阀差压	+6.017	-1.5235	13.5576
S-ZORB.FC_1203.PV	D121 去稳定塔流量	-1.3211	-4.1566	1.5143
S-ZORB.TC_1606.PV	反应器入口温度	+16.3857	8.3598	24.4116
S-ZORB.TE_5004.DACA	稳定塔顶出口温度	+2.5239	-11.011	16.0588
S-ZORB.PDT_2409.DACA	ME-115 过滤器压差	+5.1795	-2.4047	12.7636
S-ZORB.LT_3101.DACA	D-124 液位	+4.2163	2.1644	6.2681
S-ZORB.BS_AT_2402.PV	闭锁料斗氧含量	-0.3767	-2.4709	1.7176
S-ZORB.LI_9102.DACA	D-204 液位	-6.5127	-43.8617	30.8362
S-ZORB.PC_3001.DACA	D-113 压力	+0.0198	-0.0102	0.0498
S-ZORB.FT_2803.DACA	紧急氢气去 D-102 流量	+1.4543	0.7371	2.1714
S-ZORB.FC_1201.PV	D104 去稳定塔流量	-6.2666	-20.8397	8.3065

S-ZORB.FC_1202.PV	D121 顶去放火炬流量	+145.2264	65.8184	224.6344
原料性质	芳烃,v%	/不变	/	/
原料性质	辛烷值 RON	/不变	/	/

【操作优化方案】

主要变量的操作优化方案可分为以下四个部分：

（1）大幅度提升的操作变量：

闭锁料斗 H2 过滤器出口气流量、D-109 吸附剂料位、R-102 底滑阀差压、反应器入口温度、D121 顶去放火炬流量；

（2）大幅度降低的操作变量：

E-101F 管程出口管温度、D-204 液位、D104 去稳定塔流量；

（3）小幅度提升的操作变量：

闭锁料斗烃含量、精制汽油出装置温度、D104 液面、ME-109 过滤器差压、稳定塔顶出口温度、ME-115 过滤器压差、D-124 液位、D-113 压力、紧急氢气去 D-102 流量；

（4）小幅度降低的操作变量

D123 冷凝水罐液位、加热炉炉膛压力、再生器温度、D121 去稳定塔流量、闭锁料斗氧含量；

在实际操作过程中，不同原料的样本操作可进行不同的微调，主要根据表 7.1 的**置信区间**进行微调。

7.5.3 模型合理性验证

本题模型通过多次迭代，求解多目标最优解。这里列举 10 次迭代过程中，满足约束条件的样本契合度，如图所示：

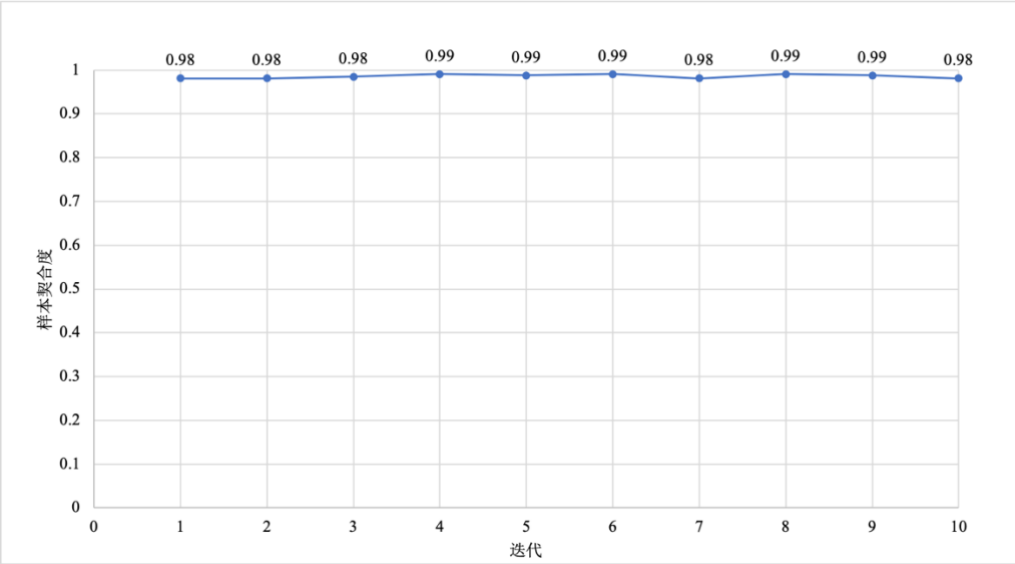


图 7.7 样本契合度曲线

多次迭代求最优解，我们采用多次实验最优解的平均值，来提供整体样本宏观上的最优操作方案。同时，进行模型合理性验证的主要观点可归纳如下：

（1）多次实验最优解的平均值，满足约束条件的样本契合度**均在 98%、99%左右**，因此**验证了模型的稳定性和合理性**；

（2）多次迭代的最优解使得所有样本中，大部分硫含量偏向接近于 **3.2 μg/g**，少部分接近于 **3.4 μg/g**；多数样本的辛烷值损失接近于 **0.4**，所有样本总体上达到了近似最优解。

7.6 小结与讨论

求解本题的过程中，我们主要根据多次实验的多次迭代求解最优值，根据每次实验的近似最优解求得代表 325 个样本整体的操作方案，使得尽可能多的样本满足硫含量和辛烷值损失降幅的具体约束。如果想要细化确定某一样本的具体方案，可以重点考虑以下几点：

（1）将样本具体划分类别，根据不同类别，收集其原料汽油精制过程中的多样本大数据信息，以供后续的数据挖掘与建模分析；

（2）汽油精制过程是一个多流程多样化的过程，可以进行不同时间段，收集具体信息，进行多分时段的建模研究。

八、模型可视化与分析

8.1 问题分析

为了催化裂化汽油精制过程的平稳、安全，需逐步调整主要操作变量，保证各变量每次按照允许调整幅度值 Δ 进行优化。本题基于第四问的优化结果，引入贪婪算法，针对每次优化调整后的结果，绘制主要操作变量优化调整过程中对应的汽油辛烷值和产品硫含量变化轨迹，实现可视化。

8.2 可视化分析

本题沿用了第四问中第 133 号样本的多次迭代最优解平均值作为增降服务的参考，使用贪心算法对主要变量进行逐步调整，展现了汽油辛烷值和硫含量的变化轨迹。

贪心算法是将调整过程分成若干个步骤，在每个步骤都应用贪心原则，选取当前状态下最好的或最优的选择（局部最有利的选择），且每次决策都以当前情况作为基础并进行选择，在每个步骤的局部最优解确定以后，就不再进行回溯处理，直到算法结束。

具体调整的步长信息如下表，总调整步长为 259。

表 8.1　操作变量调整步长的具体信息

变量代号	下界	上界	增幅	原值	迭代参考值	调整步长
S-ZORB.BS_AT_2401.PV	-0.5	5	0.1	0.0906	1.79708	**18**
S-ZORB.TE_5202.PV	30	45	1	34.4525	1.93674	**1**
S-ZORB.LC_3301.DACA	45	55	5	50.174	-1.26317	**0**
S-ZORB.PT_6002.PV	-0.6	-0.15	-0.1	-0.4361	0.03295	**0**
S-ZORB.TE_1104.DACA	50	400	1	361.9926	-180.62951	**180**
S-ZORB.FT_3702.DACA	0	60	5	0	30.01482	**6**
S-ZORB.LC_1201.PV	45	55	5	49.9801	0.61697	**0**
S-ZORB.PDT_3502.DACA	5	20	1	12.022	1.74955	**1**
S-ZORB.LT_2901.DACA	-0.05	70	5	-0.016	31.01487	**0**
S-ZORB.TC_2607.PV	450	520	1	505.0447	-6.33428	**6**
S-ZORB.PDT_2606.DACA	-0.5	20	1	10.1016	3.69219	**3**
S-ZORB.FC_1203.PV	5	15	1	11.2819	-1.57679	**1**
S-ZORB.TC_1606.PV	400	450	1	408.3889	12.20879	**12**
S-ZORB.TE_5004.DACA	40	80	1	56.9466	7.5446	**7**
S-ZORB.PDT_2409.DACA	-0.5	25	1	5.7485	6.08546	**6**
S-ZORB.LT_3101.DACA	-1.8	7	0.5	-1.2811	4.8626	**9**
S-ZORB.BS_AT_2402.PV	-0.5	5	1	0.1933	-0.08849	**0**
S-ZORB.LI_9102.DACA	10	90	5	52.7806	-12.31721	**2**
S-ZORB.PC_3001.DACA	0	0.15	0.05	0.05	0.01395	**0**
S-ZORB.FT_2803.DACA	0	3	0.5	0	1.53916	**3**
S-ZORB.FC_1201.PV	75	150	5	114.1334	-6.70444	**1**
S-ZORB.FC_1202.PV	0	300	30	43.1639	92.95539	**3**

最终结果如图所示。**不同颜色代表对不同操作变量的调整，总调整步长为 259**。在**第 234 步时达到最优，辛烷值最高为 88.7，硫含量为 3.4** μg/g。当调整步长为 **228 时，硫含量最低为 3.2** μg/g。

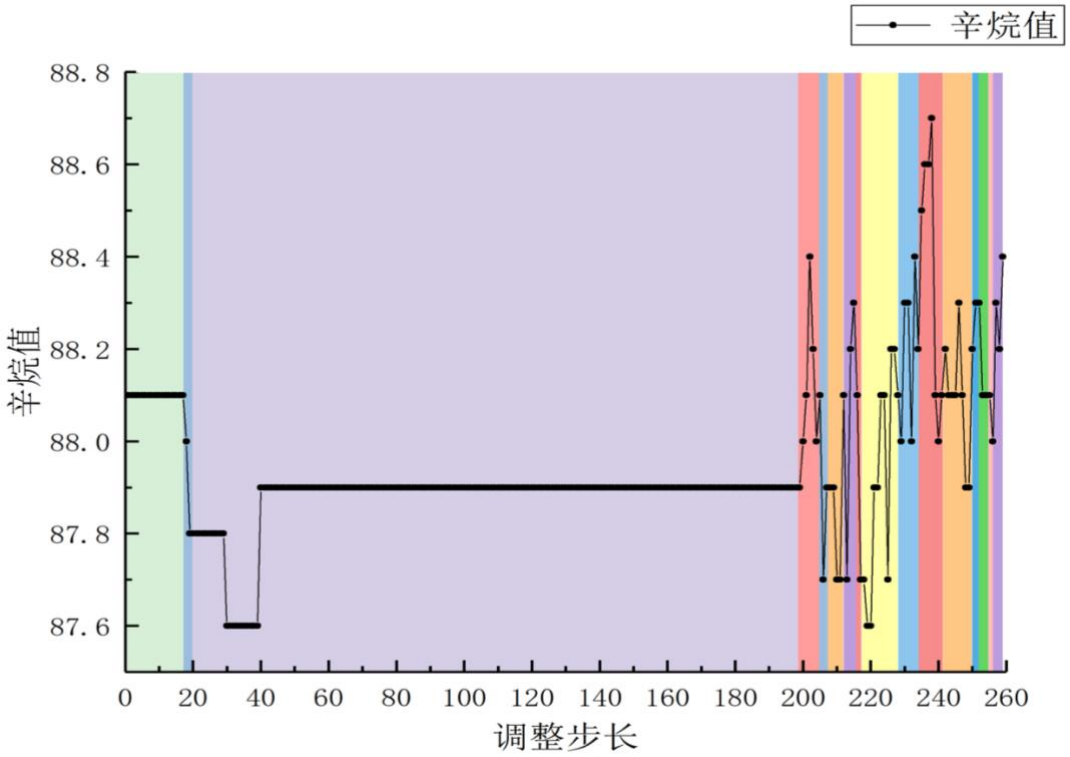


图 8.1 辛烷值变化轨迹

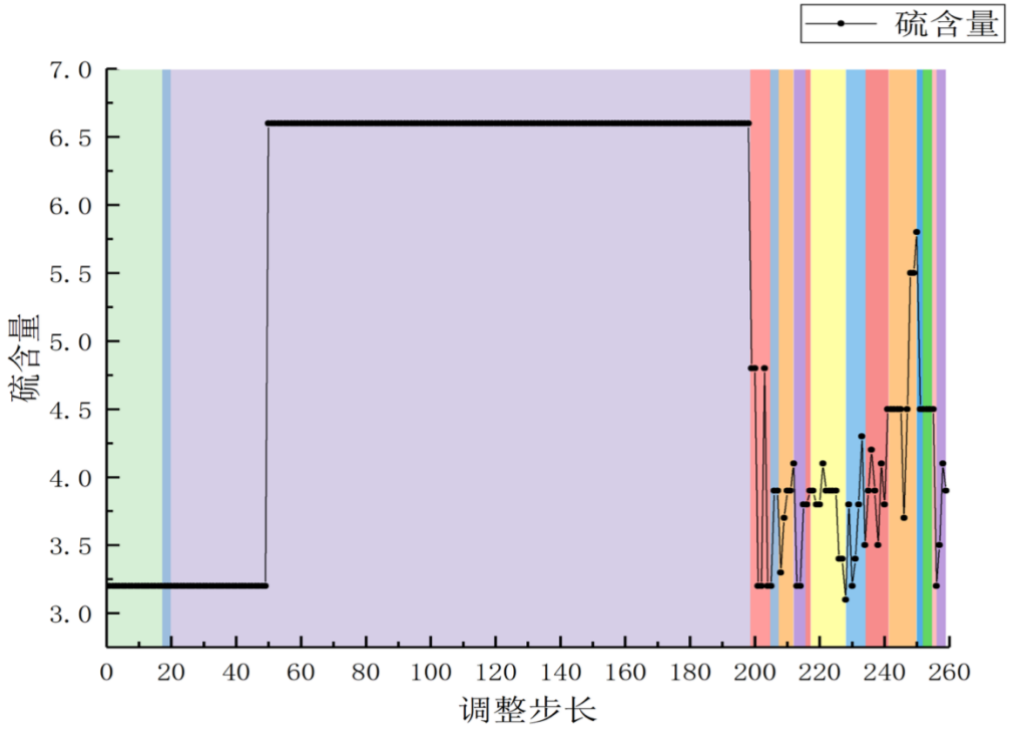


图 8.2 硫含量变化轨迹

第 234 步时达到最优，即根据表格中的操作变量顺序，以及步长进行调整。从表格中可以看出，从上至下在对**稳定塔顶出口温度**（S-ZORB.TE_5004.DACA）调整的**第 6 次**时达到最优值；在总体调整步长时，辛烷值变化、硫含量变化中存在平缓直线的区域，是由于其单位精确度为小数点后一位，具体情况是存在**微小变化**的。

参考文献

[1] 姚明海, 王娜, 齐妙等. 改进的最大相关最小冗余特征选择方法研究[J]. 计算机工程与应用, 50(9):116-122, 2014.
[2] 张延良, 卢冰. 基于信息增量特征选择的微表情识别方法[J]. 计算机工程, 45(05):261-266, 2019.
[3] Xianlun T , Nianci L , Dong Y , et al. Study on simulation of model predictive control based on RBF-SVR[J]. Experimental Technology and Management, 2016.
[4] 王重, 刘黎明. 拟合优度检验统计量的设定方法[J]. 统计与决策, (05):154-156,2010.
[5] Jie W , University W B . Comparison of the Effects of MSE and MAE on Machine Learning Performance Optimization[J]. China Computer & Communication, 2018.
[6] 王月. 最大信息系数的算法分析及改进[D]. 2020.
[7] Mundra P A , Rajapakse J C . SVM-RFE With MRMR Filter for Gene Selection[J]. IEEE Transactions on Nanobioence, 9(1):31-37, 2010.
[8] Wang Yejinpeng, Chen Liang, Ju Lingao, et al. Tumor mutational burden related classifier is predictive of response to PD-L1 blockade in locally advanced and metastatic urothelial carcinoma.. 2020, 87:106818.
[9] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版), 2014(01):142-146.
[10] Peng Hanchuan, Long Fuhui, Ding Chris. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.. 2005, 27(8):1226-38.
[11] A. C. Blanco, J. B. Babaan, J. E. Escoto, et al. MODELLING OF LAND SURFACE TEMPERATURE USING GRAY LEVEL CO-OCCURRENCE MATRIX AND RANDOM FOREST REGRESSION. 2020, XLIII-B3-2020:23-28.
[12] 龚越, 罗小芹, 王殿海, et al. 基于梯度提升回归树的城市道路行程时间预测[J]. 浙江大学学报(工学版), 2018, 52(03):46-53.
[13] 张杰, 于洋, 雷洋. 多目标博弈问题的求解算法[J]. 吉林大学学报(理学版),2016,54(04):700-708.
[14] 罗校清. 使用角度选择策略的第二代Pareto强度进化算法[J]. 计算机应用与软件, 2018, v.35(07):296-304.

附录:文件说明

代码文件说明：

代码文件名	说明
test1_1.py	提取附件三超范围的数值
test1_2.py	提取附件三超范围的数值
test1_3.py	3*detal 准则清洗异常值
test2-1.py	增量特征选择与 SVR 回归
test2-2.py	多准则验证筛选结果
test3-1.py	四种回归预测辛烷值损失
test4.py	辛烷值损失、硫含量的决策树预测
	PareTo 强度进化算法
distribution.m	分布模型拟合
dealdata.py	多次实验迭代数据处理
dealdata2.py	多次实验迭代数据处理
test5_1.py	分步提升 Δ 值

表 4.2 非契合数据整理表

变量编号	清洗后数据	附件 1 数据	契合度
10	245.2808	245.3339	99.98%
26	2950.1315	3007.9629	98.08%
33	0.9932	0.9929	99.97%
39	490.5506	555.3538	88.33%
43	**23.8203**	**44.4755**	**53.56%**
44	52.1403	53.5156	97.43%
56	0.0911	0.0902	99.02%
59	-0.0456	-0.0330	61.87%
62	413.5209	414.4614	99.77%
63	-0.1794	-0.1677	93.04%
65	0.3828	0.3831	99.93%
68	49.9602	49.9688	99.98%
72	50.0740	49.9861	99.82%
80	0.1106	0.1101	99.56%
83	498.5036	499.4726	99.81%
86	41.7503	44.8987	92.99%
87	14795000.0000	14755121.2500	99.73%
94	18300000.0000	18346556.0000	99.75%
99	21200000.0000	21155841.0000	99.79%
100	10500000.0000	10469520.0000	99.71%
104	48200000.0000	48172662.0000	99.94%
105	23700000.0000	23661272.0000	99.84%
109	49.7074	47.2221	94.74%
127	0.3837	0.3839	99.93%
130	82.3280	82.3983	99.91%
132	-1.5860	-1.5915	99.65%
133	0.4086	0.4059	99.35%
136	51.0655	51.2234	99.69%
137	0.2998	0.2974	99.19%
139	-1.2582	-1.2577	99.96%
149	0.1047	0.1047	99.94%
154	3.6521	3.6127	98.91%
156	0.1333	0.1553	85.84%
157	0.1492	0.1650	90.45%
158	-0.3451	-0.3489	98.92%
160	5.3283	5.4119	98.46%
162	**18.8566**	**33.0150**	**57.12%**
167	-4.9695	-4.9702	99.99%
175	287.5824	217.1578	67.57%
180	497.3619	498.2707	99.82%
181	497.2333	498.1257	99.82%
183	-0.5231	-0.7441	70.30%

185	-0.5388	-0.7478	72.05%
194	0.0502	0.0500	99.64%
197	379.2369	380.0332	99.79%
198	-0.0096	-0.0097	99.04%
210	1.7790	1.5117	82.32%
213	1.6416	1.4187	84.29%
215	0.3847	0.3861	99.63%
217	43.9875	42.0297	95.34%
219	0.7006	0.9027	77.61%
221	0.7025	0.9084	77.34%
227	0.1644	0.1648	99.77%
228	**6.7575**	**0.9752**	**492.93%**
229	85.6598	96.8920	88.41%
232	496.2300	497.1503	99.81%
241	0.7033	0.9054	77.68%
242	20.7407	27.2878	76.01%
243	59.1257	49.2435	79.93%
244	0.7244	1.0227	70.83%
248	2.8707	3.9831	72.07%
251	-1.8800	-1.9232	97.75%
255	545.2407	536.6471	98.40%
258	0.0062	0.0060	97.10%
260	0.0007	0.0006	86.58%
287	-0.1580	-0.1459	91.74%
288	-0.0798	-0.0635	74.44%
291	418.7470	420.2032	99.65%
292	416.2133	417.4677	99.70%
293	414.1384	414.2342	99.98%
313	0.0887	0.0718	76.38%
315	0.5713	0.5755	99.29%
316	0.4757	0.4694	98.67%
317	0.5455	0.5571	97.92%
319	0.6951	0.7039	98.75%
320	0.6523	0.6792	96.04%
322	0.6721	0.6748	99.61%
323	0.2472	0.2419	97.81%
330	49.7074	47.2221	94.74%
333	82.3255	82.3965	99.91%
348	87900000.0000	87939278.2500	99.96%

附录“6-1 回归预测结果”

支持向量回归	决策树回归	随机森林回归	梯度提升树回归
1.3	1.4	1.4	1.4
1.3	1.1	1.2	1.2
1.3	1.4	1.4	1.3
1.3	1.4	1.4	1.3
1.3	1.4	1.3	1.3
1.3	1.4	1.4	1.4
1.3	1.2	1.3	1.3
1.3	1.3	1.3	1.4
1.2	1.1	1.2	1.3
1.3	1.4	1.4	1.4
1.3	1.3	1.3	1.2
1.3	1.3	1.3	1.3
1.4	1.5	1.4	1.4
1.3	1.4	1.4	1.4
1.6	1.7	1.6	1.6
1.6	1.6	1.6	1.6
1.2	1.1	1.3	1.3
1.4	1.6	1.4	1.4
1.2	1.1	1.2	1.2
1.3	1.2	1.2	1.2
1.4	1.4	1.4	1.3
1.1	1	1.1	1
1.5	1.6	1.5	1.4
1.3	1.4	1.3	1.3
1.3	1.4	1.4	1.4
1.4	1.5	1.4	1.4
1.3	1.4	1.4	1.4
1.2	1.1	1.2	1.2
1.3	1.4	1.4	1.4
1.3	1.2	1.3	1.3
1.6	1.7	1.6	1.5
1.3	1.4	1.3	1.3
1.2	1.1	1.1	1.2
1	0.9	1	1
1.2	1.2	1.2	1.2
1.3	1.4	1.3	1.3
1.3	1.2	1.3	1.3
1.3	1.3	1.3	1.3
1.3	1.2	1.2	1.2
1.3	1.4	1.3	1.3
1.4	1.5	1.4	1.4
1.3	1.2	1.2	1.3

1.4	1.4	1.4	1.4
1.3	1.4	1.3	1.3
1.3	1.2	1.2	1.2
1.3	1.2	1.2	1.2
1.3	1.2	1.3	1.3
1.3	1.2	1.2	1.3
1.2	1.1	1.1	1.1
1.1	1	1.1	1.1
1.3	1.2	1.2	1.3
1.3	1.4	1.4	1.3
1.3	1.2	1.2	1.2
1.3	1.2	1.3	1.4
1.3	1.2	1.2	1.3
1.3	1.4	1.3	1.4
1.2	1.1	1.2	1.2
1.3	1.3	1.3	1.2
1.7	1.8	1.6	1.6
1.3	1.2	1.3	1.4
1.6	1.8	1.6	1.7
1.7	1.7	1.7	1.6
1.5	1.7	1.6	1.6
1.5	1.6	1.5	1.4
1.2	1.2	1.1	1.2
1.3	1.3	1.4	1.4
1.3	1.2	1.3	1.3
1.3	1.4	1.5	1.4
1.5	1.6	1.6	1.5
1.7	1.7	1.7	1.7
1.7	1.8	1.7	1.6
1.4	1.5	1.4	1.5
1.3	1.4	1.4	1.4
1.3	1.2	1.4	1.4
1.3	1.4	1.4	1.4
1.7	1.8	1.7	1.6
1.4	1.5	1.5	1.5
1.3	1.6	1.5	1.4
1.6	1.6	1.6	1.5
1.3	1.2	1.3	1.4
1.1	1.1	1.2	1.2
1.5	1.6	1.5	1.5
1.5	1.6	1.5	1.6
1.7	1.6	1.7	1.6
1.3	1.1	1.3	1.3
1.4	1.6	1.4	1.4

1.1	1	1.2	1.1
1.3	1.2	1.2	1.2
1.3	1.2	1.2	1.2
1.1	0.9	1.1	1.1
1.5	1.6	1.5	1.5
1.2	1.1	1.2	1.2
1.3	1.4	1.4	1.3
1.3	1.2	1.3	1.3
1.3	1.2	1.3	1.2
1.2	1.1	1.2	1.2
1.3	1.2	1.2	1.3
1.3	1.2	1.2	1.3
1.3	1.1	1.2	1.2
1.3	1.4	1.3	1.2
1.7	1.8	1.7	1.7
1.3	1.2	1.2	1.3
1.3	1.3	1.3	1.3
1.2	1.1	1.2	1.2
1.3	1.4	1.3	1.3
1.3	1.3	1.3	1.3
1.6	1.5	1.6	1.6
1.6	1.8	1.6	1.5
1.3	1.3	1.3	1.3
1.3	1.4	1.4	1.4
1.3	1.2	1.2	1.3
1.3	1.4	1.3	1.3
1.3	1.2	1.3	1.3
1.5	1.6	1.5	1.6
1.3	1.3	1.3	1.3
1.3	1.1	1.2	1.3
1.3	1.2	1.2	1.3
1.3	1.2	1.2	1.2
1.2	1.1	1.2	1.2
1.2	1.1	1.2	1.2
1.5	1.6	1.5	1.4
1.2	1.1	1.1	1.2
1.3	1.2	1.2	1.3
1.2	1.2	1.2	1.3
1.3	1.3	1.3	1.3
1.2	1.2	1.2	1.2
1.2	1.1	1.2	1.3
1.2	1.1	1.2	1.3
1	0.8	1	1.1
1.3	1.4	1.4	1.3

1.5	1.7	1.4	1.4
1.4	1.5	1.5	1.4
1.3	1.2	1.3	1.3
1.3	1.3	1.3	1.3
1.7	1.8	1.6	1.6
1.4	1.5	1.4	1.4
1.4	1.6	1.5	1.4
1.6	1.7	1.6	1.4
0.9	0.9	1	1.1
1.3	1.4	1.4	1.3
1.1	1.1	1.1	1.1
0.3	0.6	0.6	0.3
1.4	1.5	1.4	1.4
1.4	1.4	1.4	1.4
1.3	1.2	1.3	1.3
1.4	1.5	1.4	1.4
1.3	1.2	1.3	1.3
1.3	1.5	1.3	1.3
1.7	1.8	1.6	1.5
1.2	1.1	1.1	1.3
1.7	1.8	1.6	1.6
1.3	1.3	1.3	1.3
1.7	1.8	1.6	1.7
1.3	1.3	1.3	1.4
1.3	1.4	1.4	1.4
1.3	1.4	1.3	1.3
1.3	1.4	1.4	1.3
1.2	1	1.2	1.2
1.2	1.2	1.2	1.2
1.1	1	1.1	1.1
1.3	1.4	1.2	1.3
1.3	1.2	1.2	1.2
1.2	1	1.1	1.1
1.1	1	1	1.1
1.1	1	1.1	1.1
1.4	1.4	1.4	1.3
0.7	0.8	0.8	0.7
1.1	1	1	1
1.2	1.1	1.1	1.1
1.1	1	1	1
1	0.9	1	1
1	0.8	0.9	0.9
1.6	1.8	1.5	1.5
1.3	1.2	1.2	1.2

1.1	1	1	1.1
1.2	1.2	1.1	1.2
1.1	1	1	1.1
1.2	1.1	1.1	1.1
0.6	0.5	0.7	0.6
1.3	1.2	1.2	1.1
1.3	1.2	1.1	1.1
1.3	1.4	1.3	1.2
1.1	1	1	1.1
1.1	1	1	1.1
0.5	0.5	0.7	0.5
1.2	1.1	1.1	1.1
1.2	1.1	1	1.1
1.2	1.1	1.1	1.1
1.3	1.4	1.3	1.3
1	0.6	1	1
1.3	1.2	1.3	1.3
1.4	1.5	1.4	1.4
1.4	1.5	1.4	1.4
0.8	0.8	0.9	0.8
0.8	0.6	0.9	0.8
1.3	1.2	1.2	1.2
1.1	1.1	1.1	1.2
1.3	1.4	1.4	1.4
1.2	1.1	1.1	1.1
1.3	1.3	1.3	1.2
1.3	1.4	1.4	1.4
1.3	1.2	1.2	1.3
1.3	1.3	1.2	1.2
1.3	1.4	1.4	1.3
1.3	1.4	1.4	1.4
1.3	1.2	1.3	1.3
1.2	1.1	1.1	1.2
1.3	1.4	1.3	1.2
1.1	1	1.1	1.2
1.1	0.8	1.1	1.1
1.2	1.1	1.1	1.2
1.2	1.1	1.1	1.1
1.2	1.1	1.1	1.1
1.2	1.1	1.1	1.1
1.3	1.2	1.2	1.2
1.3	1.2	1.2	1.2
1.3	1.3	1.3	1.2
1.3	1.3	1.2	1.2

1.2	1	1.1	1.1
1	1	1	1
1.3	1.2	1.2	1.2
1.3	1.4	1.3	1.3
1.3	1.4	1.3	1.3
1.1	0.9	1.1	1.1
1.3	1.3	1.2	1.3
1.3	1.3	1.3	1.3
1.3	1.3	1.3	1.2
1.3	1.3	1.3	1.3
1.3	1.3	1.3	1.3
1.3	1.3	1.3	1.2
1.2	1.1	1.1	1.1
1.2	1.3	1.3	1.2
1.2	1.3	1.3	1.2
1.2	1.2	1.2	1.2
1.2	1.1	1.1	1.2
1.2	1.1	1.1	1.2
1.2	1.1	1.1	1.2
1.2	1.2	1.2	1.2
1.3	1.2	1.2	1.2
1.2	1.2	1.3	1.2
1.1	0.9	1	1.1
1	0.9	0.8	0.9
1	0.9	1	1.1
1.2	1.1	1.1	1.2
1.3	1.3	1.2	1.2
1.2	1.1	1.1	1.2
1.3	1.3	1.3	1.2
1.3	1.1	1.2	1.2
1.3	1.3	1.3	1.3
1.3	1.2	1.2	1.2
1.2	1.1	1.1	1.2
1.3	1.3	1.3	1.2
1.3	1.3	1.3	1.2
1.3	1.2	1.2	1.2
1.3	1.2	1.2	1.2
1.3	1.3	1.3	1.2
1.3	1.2	1.2	1.2
1.3	1.1	1.2	1.2
1.3	1.2	1.2	1.2
1.3	1.2	1.2	1.2
1.3	1.4	1.3	1.2
1.3	1.2	1.2	1.2

1.3	1.2	1.2	1.2
1.3	1.4	1.2	1.3
0.8	0.8	0.8	0.8
1.3	1.2	1.1	1.2
1.3	1.2	1.2	1.1
1.3	1.3	1.2	1.2
1.3	1.3	1.3	1.2
0.9	0.9	1	1
1.3	1.2	1.3	1.2
1.3	1.3	1.2	1.2
1.3	1.2	1.2	1.2
1.3	1.3	1.3	1.2
1.3	1.3	1.3	1.3
1.2	1.1	1.1	1.2
1.3	1.3	1.3	1.3
1.2	1.1	1.1	1.1
1.2	1.1	1.1	1.2
1.3	1.1	1.2	1.2
1.2	1.1	1.1	1.2
1.2	1.1	1.1	1.1
1.1	1.1	1	1.1
1.2	1.1	1.1	1.1
1.2	1.1	1.1	1.2
1.3	1.3	1.2	1.2
1.2	1.1	1.1	1.1
1.2	1.2	1.2	1.1
1.3	1.2	1.2	1.3
1.2	1.1	1.1	1.1
1.3	1.1	1.2	1.2
0.7	0.5	0.8	0.7
1.2	1.1	1.1	1.2
1.2	1.1	1.1	1.2
1.3	1.3	1.3	1.2
1.2	1.1	1.1	1.1
1.2	1.1	1.1	1.1
1.3	1.2	1.2	1.2
1.2	1.1	1.1	1.2
1.3	1.2	1.2	1.2
1.2	1.1	1.1	1.2
1.3	1.5	1.3	1.3
1.5	1.5	1.4	1.4
1.2	1.1	1.1	1.2
1.3	1.2	1.2	1.2
1.2	1.1	1.1	1.2

1.1	0.9	1.1	1.1
1.3	1.3	1.2	1.2
1.2	1.1	1.1	1.1
1.3	1.4	1.3	1.3
1.3	1.2	1.2	1.2
1.2	1.1	1.1	1.1
1.3	1.2	1.2	1.2
1.3	1.2	1.2	1.2
1.3	1.2	1.2	1.2
1.2	1.1	1.1	1.2
1.2	1.1	1.1	1.2
1.3	1.3	1.3	1.3
1.2	1.1	1.2	1.2
1.3	1.3	1.3	1.3
1.2	1.1	1.1	1.1
1.3	1.2	1.3	1.3
1.3	1.3	1.3	1.2
1.3	1.3	1.3	1.2
1.3	1.3	1.3	1.3